Next Generation Sequencing Approaches to Identify Novel Susceptibility Genes for Epithelial Ovarian Cancer

Jane D. Hayward

University College London

PhD

2014

Declaration

I, Jane D. Hayward, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Ovarian cancer is the fifth most common cancer in women in developed countries and is associated with poor survival due to late diagnoses. Strategies focusing on detecting the disease in the earliest stages and/or improving risk prediction may represent effective clinical intervention reducing disease burden. Women at the greatest risk of epithelial ovarian cancer (EOC) can be offered prophylactic risk-reducing salpingo-oopherectomy (RRSO), which is currently only offered to women with mutations in the highly penetrant susceptibility genes *BRCA1* or *BRCA2*. Previous studies show that 46% of familial cases of EOC carry a deleterious mutation in *BRCA1* (37%) or *BRCA2* (9%). The residual proportion of familial risk is likely to be attributable to other genetic variants providing a rationale for identifying additional susceptibility alleles using rapid high-throughput next generation sequencing (NGS) in large samples sizes.

A pilot study determines the principle of NGS in mutation detection sequencing *BRCA1* gene in 12 DNA samples with known mutations. The 11bp deletion, missed in the analysis, is detected by altering the bioinformatics. The second study sequences 1506 cases and 1130 healthy controls using Fluidigm microfluidic technology and Illumina HiSeq2000 in 6 DNA repair genes (*RAD51B*, *RAD51C*, *RAD51D*, *XRCC2*, *XRCC3* and *SLX4*). 94% of the targeted region is sequenced with >30 reads. 23 cases and 1 control show a putative protein-truncating variant in 5 genes. Many missense variants are detected in cases and controls suggesting these are not pathogenic. Epidemiological data shows that women with family history and a deleterious mutation develop EOC on average 10 years younger. Interestingly, half of those women with detected mutations have no family history.

A final study uses the established NGS approach to characterise the mutation prevalence in 4 known and 5 candidate EOC susceptibility genes in 2300 unaffected women from high-risk breast-ovarian cancer families. *BRCA1* and *BRCA2* deleterious mutations are identified in 53 and 49 women respectively. Deleterious mutations are detected in 6 additional genes, *BRIP1* (n=5), *RAD51C* (n=3), *RAD51D* (n=1) *PALB2* (n=5), *BARD1* (n=1) and *NBN* (n=3). Importantly, a bioinformatics pipeline is refined to maximise variant detection sensitivity with zero false negatives where read depth is >30X. Further large case-control studies are recommended to examine the population frequencies in these novel genes. These studies demonstrate the potential of targeted NGS approaches for population-wide risk prediction and early detection of EOC.

Table of Contents

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

List of Tables

List of Figures

Acknowledgements

There are a number of people that I would like to thank for their support. I would like to thank my Supervisors, Prof. Simon Gayther and Dr Susan Ramus for giving me the opportunity to undertake this PhD and for their supervision over the last few years. I would also like to thank all of the scientists at the Source Bioscience sequencing laboratory in Nottingham, especially Dr Tom Burr and Dr Cliff Murray, Mr Simon Mayes and Dr Barry Murphy.

I thank Dr Ed Dicks for Bioinformatics support and Dr Honglin Song, both from Strangeways Research Laboratory, Cambridge. I would also like to thank the team at the University of Southern California, Maria Intermaggio, Andre Kim and Christopher K Eland.

I thank all the scientists at the Great Ormond Street Molecular Genetics Laboratory for allowing me to work in their laboratory and use their Fluidigm equipment, especially Dr Angela Barrett and Dr Suzanne Drury.

I would like to extend a particular thank you to Dr Sioban SenGupta, whom has been my Graduate Tutor throughout much of my time at the Institute for Women's Health. Your consistent support and encouragement have been pivotal in my completion of this thesis. I thank all of my colleagues and friends in the Women's Cancer Department; you have all made the Department one that I am proud to have been a part of; with particular thanks to Prof. Martin Widschwendter, Ms Allie Jones, Ms Eva Wozniak, Dr Shahzia Anjum, Dr Jacqueline McDermott, Dr Madhuri Salker, Ms Sue Philpott and Dr Andy Ryan.

I am indebted to my friend, Jane Cameron, for her enthusiasm in proofreading this thesis. Thank you, Jane.

Finally, I thank my Mother, Monica, for your inspiration; and I thank my Father, Peter, for your unwavering support and encouragement.

List of Abbreviations

Abbreviation	Definition
AOCS	Australian ovarian cancer study
ATM	Ataxia telangiectasia mutated
ATR	ATR serine/threonine kinase
BAM	Binary alignment map
BARD1	BRCA1 associated RING domain 1
BIC	Breast cancer information core
bp	base pairs
BRAF	B-Raf proto-oncogene, serine/threonine kinase
BRCA1	Breast cancer 1 early onset
BRCA2	Breast cancer 2 early onset
BRIP1	BRCA1 interacting protein C-terminal helicase 1
BWA	Burrows-Wheeler Aligner
CA-125	cancer antigen 125
CHK2	Checkpoint kinase 2
CIMBA	Consortium of Investigators of Modifiers of BRCA1 and BRCA2
CS1	custom adapter sequence tag 1
CS2	custom adapter sequence tag 2
dbSNP	database of short genetic variation
Del	deletion
EOC	Epithelial ovarian cancer
ERRB2	v-erb-b2 avian erythroblastic leukaemia viral oncogene homolog 2
FA	Fanconi anaemia
FAP	Familial Adenomatous Polyposis
FDR	First degree relative
FH	Family history
FIGO	International Federation of Gynaecology and Obstetrics
FS	Frameshift mutation
GATK	Genome Analysis Toolkit
Gb	gigabase pairs
GRFOCR	Gilder Radner familial ovarian cancer registry
GWAS	Genome wide association study
HBOC	Hereditary breast and ovarian cancer syndrome
HGSOC	High grade serous ovarian carcinoma
HNPCC	Hereditary nonpolyposis colorectal cancer
HR	homologous recombination
IFC	integrated fluidic circuit
IGV	Integrative genomics viewer
indel	insertion/deletion
Ins	insertion
Kb	kilobase pairs
KRAS	Kirsten rat sarcoma viral oncogene homolog
MAF	minor allele frequency

MALOVA	Malignant ovarian cancer study
Mb	megabase pairs
MLPA	multiplex ligation probe amplification
MRE11	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)
MS	Missense mutation
NBN	Nibrin
NGS	Next Generation Sequencing
NHEJ	non-homologous end joining
NICE	National Institute for Health Care and Excellence
NS	Nonsense mutation
NS-SNV	non-synonymous single nucleotide variant
PALB2	partner and localizer of BRCA2
PARP	Poly (ADP-ribose) polymerase
PCR	Polymerase chain reaction
PE1	paired end sequence primer 1
PE2	paired end sequence primer 2
PGD	preimplantation genetic diagnosis
PROMISE	Predicting Risk of Ovarian Malignancies, Improved Screening and Early detection
PTEN	Phosphatase and tensin homolog
QC	Quality control
RAD50	RAD50 homolog (S. cerevisiae)
RAD51B	RAD51 paralog B
RAD51C	RAD51 paralog C
RAD51D	RAD51 paralog D
RRSO	Risk reducing salpingo-oopherectomy
SAM	Sequence alignment map
SDR	Second degree relative
SLX4	SLX4 structure-specific endonuclease subunit
SSA	single-strand annealing pathway
S-SNV	synonymous single nucleotide variant
TCGA	The cancer genome atlas
TP53	Tumour protein p53
TS	target specific primer
UKFOCR	UK familial ovarian cancer registry
UKFOCSS	UK familial ovarian cancer screening study
UKOPS	UK ovarian cancer population study
XRCC2	X-ray repair complementing defective repair in Chinese hamster cells 2
XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3

Chapter One

Introduction

1.1 The Genetic susceptibility to cancer

Research into genetic susceptibility and cancer has been a highly valuable objective within cancer research for several decades. The discovery of cancer susceptibility syndromes that result from a genetic predisposition to a variety of cancer types has guided research into novel treatment targets and provided greater insight into the biological mechanisms in tumour development (Fletcher & Houlston 2010).

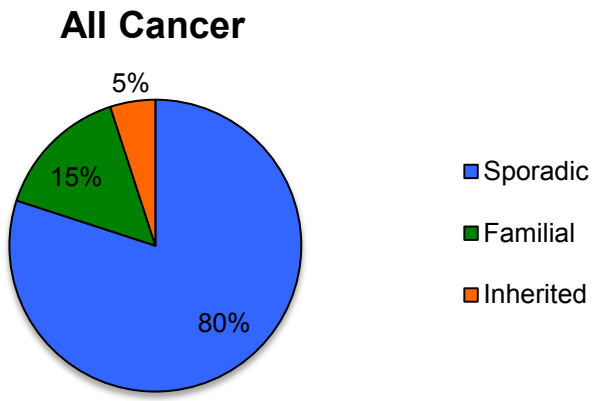


Figure 1.1 Susceptibility to all cancers
Around 80% of cancers are likely to be sporadic, with 15-20% being familial (i.e. those cancers caused by a combination of low penetrance variants and environmental factors) and those that are truly inherited (i.e. caused by rare inherited high penetrance variants) are likely to be around 5-10% of all cancer (Adapted and re-drawn from Nagy et al 2004; Oncogene (2004) 23, 6445–6470).

Cancers that occur sporadically account for the majority of diagnoses (~80%). These randomly occurring cancers include those without a family link or gene alterations that result in an elevated cancer risk. Around 15-20% of cancers are considered familial, i.e. they result from a combination of low penetrance variants and environmental factors. Cancers that are truly inherited (i.e. they are caused by rare moderate-high penetrance gene variants) account for the smallest proportion of around 5-10% and often exhibit an autosomal dominant pattern of inheritance.

1.2 Cancer susceptibility models

The earliest model of cancer susceptibility dates back to the end of the 1960s with the idea that cancers clustering in families are due to genetic changes inherited between

relatives that result in a carcinogenic 'hit' that increases the risk of cancer developing. Ashley (1969) examined colon cancer in patients with the hereditary condition polyposis coli proposing that those patients positive for the polyposis coli gene were predisposed to develop colon cancer as they already had one 'hit'. Thus, they required one or two fewer hits to develop the disease. Ashley based this conclusion on examination of the ratio of age-specific incidence of colon cancer patients in the general population against the age-specific prevalence in patients positive for the polyposis coli gene. Ashley's (1969) data showed these ratios reducing as the patient's age increased. The 'two-hit' model is now considered an early milestone in cancer research with this model forming the basis of some autosomal dominant cancer syndromes, including familial adenomatous polyposis (FAP).

1.2.1 High penetrance – rare variant model

High-penetrant cancer susceptibility genes lead to a high risk of developing cancer. The predominant high-penetrant genes, discovered from the end of the 1980s through to the early 1990s mostly use linkage analysis and positional cloning. The genes discovered at this time are breast cancer 1, early onset (*BRCA1*) and breast cancer 2, early onset (*BRCA2*) (for breast and ovarian cancer), adenomatous polyposis coli (*APC)*, *MLH1* and *MSH2* (for colorectal cancer) and cyclin-dependent kinase inhibitor 2A (*CDKN2A)* (for malignant melanoma). Continued attempts to detect other high-risk genes for colorectal cancer and for breast cancer have not been fruitful. Therefore, it is likely that the familial risk that is not attributable to the existing high-risk genes could be due to the inheritance of multiple variants with a moderate increase in risk (Fletcher & Houlston 2010). For breast cancer, deleterious variants in *BRCA1*, *BRCA2* and *TP53* occur in less than 0.5% of the population (Gayther 2012), but can increase the risk of developing the disease 10-20 fold compared to the risk for the general population without these mutations (Stratton & Rahman 2008).

1.2.2 Moderate penetrance – rare variant model

This polygenic model of cancer susceptibility proposes that several rare variants exist that confer a more moderate increase in cancer risk. Rare variants in this model include SNPs (Single Nucleotide Polymorphisms with minor allele frequencies of ≥1% but, <5%), sub-polymorphic variants (SNPs with minor allele frequencies of ≤1%) and deleterious mutations, such as insertions and deletions ranging from 1 base to many bases. Such susceptibility variants are detected in breast cancer using DNA

sequencing technologies in which candidate genes are examined in studies comparing cases with healthy controls. In breast cancer susceptibility, rare variants in this category include ataxia telangiectasia mutated (*ATM*), checkpoint kinase 2 (*CHK2*), BRCA1 interacting protein C-terminal helicase 1 (*BRIP1*) and partner and localiser of BRCA2 (*PALB2*), occurring in the population at rate of <0.6%; these rare variants are estimated to result in an increased risk of breast cancer of 2-4 times that of the general population (Stratton & Rahman 2008).

1.2.3 Low penetrance – common variant model

The low penetrance – common variant susceptibility alleles are also included as part of the polygenic model. This model suggests several commonly occurring (with minor allele frequencies >5%) SNPs associate together to increase the risk of cancer by a small extent, perhaps up to 3 times an increase compared to the general population. The predominant method employed in this model is the Genome Wide Association Study (GWAS). These genetic association studies, which examine vast numbers of polymorphic variants distributed throughout the genome, are used to identify loci for cancer susceptibility (i.e. specific chromosomal regions that are linked to an increased risk in cancer) (Gayther & Pharoah 2010). Tag SNPs are located in genomic regions of high linkage disequilibrium (that is alleles that are non-randomly linked at more than two genomic loci) and are useful in GWAS to identify these cancer susceptibility loci. SNPs associated with breast cancer, which are found in the population at a rate of 5-10%, are discovered to increase breast cancer risk by around 1.25 x that of the rest of the population with differences being noted in those individuals with homozygous or heterozygous SNPs (Stratton & Rahman 2008).

1.3 Familial cancer syndromes

Mutated tumour suppressor genes (TSG), genome stability genes or (rarely) oncogenes are the cause of familial cancer syndromes. Familial cancer syndromes often follow an autosomal dominant, but sometimes a recessive pattern of inheritance and exhibit complete or incomplete penetrance. Familial cancer syndromes frequently exhibit specific characteristics that indicate their distinction from sporadic cancers. These observations include: more than one primary tumour in the same anatomical region, for example, contralateral breast cancer, earlier onset disease, rare histological tumour subtypes, and affected first degree relative(s) (FDR) with similar characteristics (Weber et al 2001).

Some of the most highly penetrant syndromes include: Cowden Syndrome, familial adenomatous polyposis (FAP), hereditary nonpolyposis colorectal cancer (HNPCC) also known as Lynch Syndrome, hereditary breast-ovarian cancer syndrome, retinoblastoma, Li-Fraumeni syndrome and Peutz-Jeghers syndrome. All of these are very rare syndromes with penetrance levels of 70-100% (Nagy et al 2004).

Mutated tumour suppressor genes (TSG) result in a downregulation of gene function. Mutated oncogenes result in an up regulation of function or function when its normal counterpart would be inactive. Genome stability genes are involved in maintaining the integrity of the cell. This is achieved through DNA repair efficiency via several types of DNA repair mechanism; for example, base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR) and homologous recombination (HR) (Vogelstein & Kinzler 2004). However, in principle they act in a similar manner to classical tumour suppressor genes in that their mutations lead to loss of function.

The genetic basis of cancer pathogenesis is reasonably well established. If tumourigenesis arises due to uncontrolled cell growth and proliferation, then mutations in both oncogenes and tumour suppressor genes are necessary. Genome stability genes however, function in cell maintenance and are involved in the repair of genetic alterations in other genes, thus if genes in this class are mutated it can result in an increase in mutations in the other classes of genes. Tumour suppressor genes and genome stability genes have to be inactivated in order to result in downregulated, or non-functioning, genes so both alleles of a gene must be lost or inactivated. Thus, if an individual has a germline mutation in *BRCA1* (a stability gene) then they will be predisposed to develop tumours; but cancer will only develop if they incur a somatic mutation rendering the wild type allele also inactive. Oncogenes, however, need to be activated (upregulated); therefore, a mutation in just one allele is all that is required.

1.3.1 Tumour suppressor genes

Hereditary retinoblastoma is caused by mutations in the retinoblastoma 1 (*RB1*) gene, most of which are large or small deletions; retinoblastoma follows an autosomal dominant pattern of inheritance and almost complete penetrance of 90%.

Familial adenomatous polyposis (FAP, OMIM 175100) is a cancer syndrome causing colorectal cancer. It is caused by mutations in the *APC* gene and results in many adenomatous polyps through the colorectum. FAP follows an autosomal dominant

pattern of inheritance with penetrance greater than 95% (Nagy et al 2004). Cowden's Syndrome (CS) is also an autosomal dominant disease caused by mutations in the phosphatase and tensin homolog (*PTEN*) gene and results in a number of different clinical features often identified in the general unaffected populations for example, fibrocystic breast disease or leiomyoma. However, CS patients are also at an increased risk of developing breast, thyroid or endometrial cancers.

Li-Fraumeni syndrome (LFS, OMIM 151623) is caused by a heterozygous mutation in tumour protein 53 gene (*TP53*) and clinically causes multiple cancers including breast cancer, sarcomas, brain tumours, leukaemias and adrenocortical carcinoma. LFS follows an autosomal dominant pattern of inheritance with virtually complete penetrance that appears to vary according to gender. In males penetrance is approximately 68% and in females 93%. Females with LFS also tend to be diagnosed with cancers at a much younger age as children or young adults. Germline gene variants in *TP53* are noted in more than half of all LFS families.

Peutz-Jeghers syndrome (PJS OMIM 175200) is caused by mutations in the threonine kinase 11 gene (*STK11*). The syndrome presents as polyps through the gastrointestinal tract including, the colon, duodenum, stomach and jejunum. PJS has an autosomal dominant pattern of inheritance. The syndrome is extremely rare occurring in 0.0005% of the population with penetrance at around 100%.

1.3.2 Genome stability genes

Hereditary breast-ovarian cancer (HBOC, OMIM 113705) is largely caused by mutations in genes *BRCA1* and *BRCA2* and has an autosomal dominant pattern of inheritance with incomplete penetrance. The penetrance estimates of *BRCA1* and *BRCA2* reduce with advancing age and vary between breast and ovarian cancer. Antoniou et al (2003) estimate average breast cancer risks by the age of 70 to be between 45-65% and ovarian cancer to be between 11-39%, with *BRCA1* conferring higher cancer risks for both diseases.

Lynch syndrome (OMIM 120435), which is also known as hereditary non-polyposis colorectal cancer (HNPCC), results from heterozygous mutations in DNA mismatch repair (MMR) genes. HNPCC is caused by mutations in the gene mutS homolog 2, colon cancer nonpolyposis type 1 (*MSH2*). HNPCC is often diagnosed at an early age and colorectal cancers are more commonly located in a proximal region. The

syndrome is characterised by multiple primary cancers. Genes involved in HNPCC are postmeiotic segregation increased 2 (*PMS2*), mutL homolog 1, colon cancer nonpolyposis type 2 (*MLH1*) and (*MSH2*), mutS homolog 6 (*MSH6*).

Lynch syndrome follows an autosomal dominant pattern of inheritance. Prevalence of Lynch syndrome is estimated to be in the region of 0.05% at the general population level and around 1-3% in patients affected by endometrial or colorectal cancer. Lifetime (up to age 70) penetrance estimates of colorectal cancer are close to 100% in men with Lynch syndrome. In women, the lifetime penetrance estimates of endometrial cancer are up to 60% and colorectal cancer up to 54% by age 70 (Nagy et al 2004).

Fanconi Anaemia (FA, OMIM 227650) is a highly heterogeneous syndrome caused by a several FANC and FANC- like genes. The syndrome results in increased risks of cancers especially acute myeloid leukaemia. FA has an autosomal recessive pattern of inheritance.

1.3.3 Oncogenes

Multiple endocrine neoplasia type 2 (*MEN2*) is caused by the oncogene ret proto-oncogene (*RET)* and follows an autosomal dominant pattern of inheritance with almost complete penetrance. *MEN2* leads to conditions including medullary thyroid carcinoma (MTC), hyperparathyroidism (HPT) or parathyroid adenomas and phaeochromocytoma (PC) which are tumours involving the adrenal chromaffin cells.

1.4 The genetic susceptibility to ovarian cancer

1.4.1 Ovarian cancer epidemiology and aetiology

Ovarian cancer is the fifth most common cancer in women in developed countries. In the UK around 6,500 women are diagnosed and 4,400 women die from the disease each year meaning that ovarian cancer causes 6% of deaths in women from cancer. Globally, the figure is likely to be around 225,000 new cases diagnosed annually (2008). Extensive geographical variation is noted between different regions internationally as disease incidence rates appear to differ dramatically between developing and developed countries. Regions with the highest incidence rates are Northern, Central and Eastern Europe with Africa and regions of Asia showing the lower incidence rates. In Europe 65,000 new diagnoses were made in 2008 and 21,500 new cases diagnosed in the US in the same year (Cancer Research UK 2008 statistics).

There has been some progress in survival rates in the 30 years between the early 1970s until the early 2000s for short-term survival. In 1971 women diagnosed with ovarian cancer had a survival probability of 42% (1 years) and 21% (5 years). In 2003 these survival rates were 70% (1 year) and 41% (5 years). However, the long-term survival rates remain low, even with current improved treatments the five-year survival rate is still less than 50%. This is likely to be attributable to late diagnosis of the disease. The stage at diagnosis is key in survival rate: diagnoses at stage I disease show a five year survival rate of ~90%, compared to diagnoses made at stage IV where five year survival rate is as low as 5%.

Essential in improving these bleak statistics are the introduction of superior approaches to risk prediction and earlier detection. In addition, risk prediction will improve survival rates as identifying women at increased risk can be offered risk-reducing surgery known as risk-reducing salpingo-oophorectomy (RRSO). This is currently offered to women with mutations in *BRCA1* or *BRCA2* and is a successful prophylactic in these cases reducing the risk of developing ovarian carcinomas.

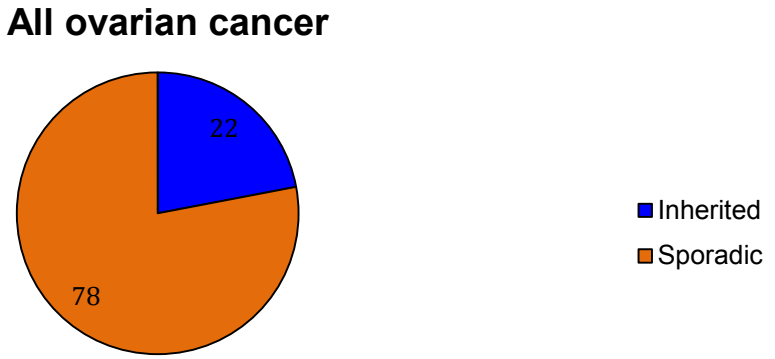


Figure 1.2 A graphical representation of the proportion of sporadic (78%) and inherited (22%) causes of ovarian cancer*. This pie chart demonstrates the large proportion of sporadic ovarian cancers compared to the small proportion of inherited ovarian cancers.*

Studies in twins estimate that around 78% of ovarian cancer is sporadic, with the remaining 22% being due to inherited genetics (Figure 1.2) (Lichtenstein et al 2000). An individual with a first degree relative (FDR) has a 3-fold increase in risk of developing ovarian cancer. Ramus et al (2007) show that in cases where there are more than two familial cases of epithelial ovarian cancer (EOC) in either first or second degree relatives, 46% are found to have probable deleterious mutations in *BRCA1* (37%) or *BRCA2* (9%).

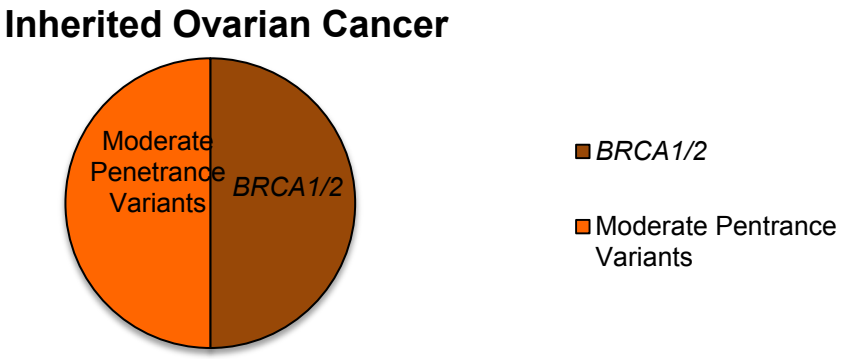


Figure 1.3 A graphical representation of the residual unknown proportion of ovarian cancer susceptibility. *This pie chart demonstrates that half of inherited ovarian cancers are attributable to moderate penetrance variants other than BRCA1 or BRCA2.*

BRCA1 and *BRCA2* have previously been shown to be the main (high penetrant) ovarian cancer susceptibility genes, accounting for ~50% of ovarian cancers. Hoskings et al (2011) suggests that in breast cancer common low penetrant variants are likely to account for ~8% of familial risk; it is plausible then that a similar contribution can be estimated in ovarian cancer. Therefore, the large residual proportion of inherited ovarian cancer risk (around 46%) may be attributable to additional rare gene variants

(Figure 1.3). This provides a strong rationale for researching additional moderate to high penetrant gene variants in epithelial ovarian cancer (EOC). These gene variants can be discovered via large scale sequencing studies of candidate genes or exome sequencing.

Additional rare variants involved in ovarian cancer may follow a similar pattern to those identified in breast cancer; in which these rare variants can be subdivided into high-penetrance and moderate-penetrance variants. Investigating the *BRCA1* and *BRCA2* molecular pathways lead to the identification of additional rare variants of moderate penetrance, namely *ATM*, *CHK2*, *BRIP1*, *PALB2* and Nijmegen breakage syndrome 1 (*NBS1*). These variants are all identified as implicated in DNA repair and linked to the *BRCA1* and *BRCA2* network. These variants have an estimated frequency of ≤0.6% (Stratton & Rahman 2008). To study the contribution of mutations in these moderate-penetrance genes to ovarian cancer would require extremely large cohorts since not only are these variants very rare they confer a lower risk of cancer; as such finding affected women would be problematical. However, this could be achieved via international collaboration initiatives (Stratton & Rahman 2008).

Figure 1.4 Allele frequency and breast cancer risk

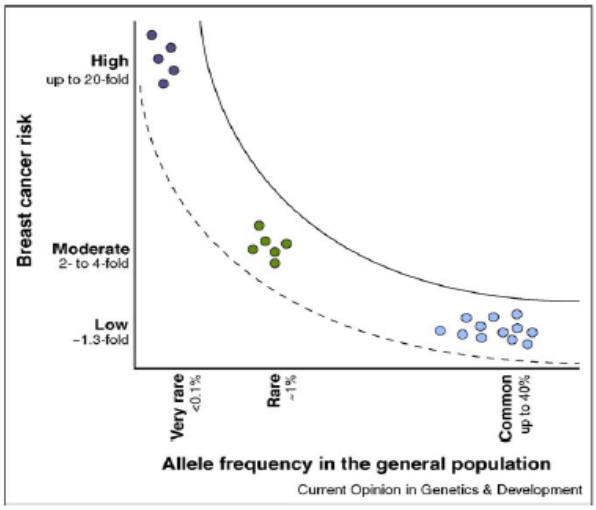


Figure 1.4 Allele frequency and breast cancer risk. *This diagram illustrates the relationship between breast cancer risk and allele frequency in the general population adapted from: Hollestelle et al Current Opinion in Genetics & Development 2010,* ***20****:268–276*

Figure 1.4 demonstrates the association between genes conferring high, low and moderate cancer risk and the allelic frequency amongst the population. High penetrance genes are rare within the population, whereas low penetrant genes are more common. Moderate risk genes are rare in occurrence (Hollestelle et al 2010).

Hollestelle et al (2010) suggest that the high penetrant genes *BRCA1* and *BRCA2* occur at a frequency of <0.1% in the population, where other estimates suggest <0.5% (Gayther 2012). It is likely that the genetic predisposition to ovarian cancer follows a similar pattern to breast cancer and that whilst the high penetrant rare alleles have been discovered through genetic linkage analysis; the moderate risk rare variants are probably best identified using DNA sequencing of large cohorts of cancer cases and controls (Li & Leal 2009).

1.4.2 Family history

Ramus & Gayther (2009) review the frequency of *BRCA1* and *BRCA2* mutations detected in breast and ovarian cancer or ovarian cancer families comparing a number of different studies conducted around the world. They report a disparity in frequencies between studies. These frequencies range from 34% to 84% for the presence of a mutation in either gene. This apparent disparity may be due to a number of issues including, mutation detection methods, cohort recruitment biases (i.e. some studies may include only those families with >2 ovarian cancer cases whilst others >1 case was sufficient to be included), the presence of founder mutation cohorts, differences in protocols adopted by laboratories, size of cohort and finally, some studies screen all possible mutations where others do not include large genomic rearrangements. Ramus & Gayther (2009) note that within ovarian cancer families mutation frequencies vary between two cancer syndromes; hereditary breast and ovarian cancer (HBOC) syndrome and site-specific ovarian cancer (ovarian cancer only). Detected mutation frequencies are higher in HBOC families (81%) compared to ovarian cancer only families (63%). These data are derived from two familial ovarian cancer registries (Ramus & Gayther 2009). In addition, these data also demonstrate that when there are more ovarian cancer cases within a family the frequency by which mutations are detected also increases. Finally, Ramus & Gayther (2009) note that the proportion of *BRCA1* mutations in families with ovarian cancer cases is higher than *BRCA2* mutations detected in these families.

1.4.3 High penetrance genes

BRCA1 and *BRCA2* are likely to be responsible for around 13% of high-grade serous ovarian adenocarcinomas (The Cancer Genome Atlas 2011, Risch et al 2006). Hereditary breast/ovarian cancer syndrome (HBOC) is probably one of the most widely documented hereditary cancer syndromes. Within the general population prevalence

figures are estimated to be in the region of 1:800 for a *BRCA1* mutation and 1:500 for *BRCA2* mutations (Antoniou et al 2008). These figures vary extensively from study to study as the prevalence statistics vary between populations examined. Inherited mutations in *BRCA1/2* genes also increase the likelihood of other cancers developing, for example cancers of the pancreas, fallopian tube, stomach, larynx and prostate (Nagy et al 2004). Penetrance figures also fluctuate between populations, which may be the result of additional altering environmental factors or additional gene variants or due to variations in risk between different gene variants (Nagy et al 2004). The penetrance rate in high-risk families (i.e. several cases within the family) is estimated to be 70% breast cancer by age 70 in those with *BRCA1* or *BRCA2* mutations. Antoniou et al (2003) calculate accurate figures for the risk of developing breast or ovarian cancer with an inherited mutation in either gene. In this study, they collect data from 22 different studies, which include 8139 cancer index cases concluding that the cumulative EOC risk for women with mutations in *BRCA1* is between 44% and 63% by 70 years and in those with mutations in *BRCA2* is between 27% and 31% by 70 years.

1.4.4 Mismatch Repair genes in Lynch syndrome (HNPCC)

A significant cause of inherited ovarian cancer is the familial cancer syndrome Hereditary Nonpolyposis Colorectal Cancer (HNPCC). Mutations in mismatch repair (MMR) genes may be responsible for up to 10% of inherited ovarian cancer. Barrow et al (2009) examine 121 families with Lynch syndrome calculating the cumulative lifetime incidence for women with mutations in the Lynch Syndrome genes as 32.5% for gynaecological cancers (endometrial and ovarian). The risk of endometrial cancer being much higher than ovarian cancer; in the study they quote the average cumulative ovarian cancer incidence of 6.1% by age 70. Lu & Daniels (2013) report the lifetime incidence of ovarian cancer in Lynch syndrome as between 6% and 8%.

1.4.5 *RAD51C* and *RAD51D* as ovarian cancer susceptibility genes

In the last two years two new ovarian cancer susceptibility genes have been identified. Meindl et al (2010) examine the gDNA of 480 women with HBOC finding 6 heterozygous deleterious mutations in RAD51 homolog C (*S. cerevisiae*) (*RAD51C*) occurring at a rate of 1.3% in the study population. This compares to zero mutations detected in women with breast cancer only. Meindl et al (2010) also examine the RAD51 locus in tumour tissues finding they exhibit loss of heterozygosity suggesting *RAD51C* is a tumour suppressor. Loveday et al (2011) analyse *RAD51D* for germline

mutations in 911 women from HBOC families and 1,060 matched healthy controls. They find 8 deleterious mutations in the cases and just one in the controls. Loveday et al (2011) report the link between *RAD51D* and cancer is more significant in ovarian cancer cases than breast cancer, calculating the relative risks to be 6.30 and 1.32 for ovarian and breast cancer respectively.

1.4.6 Genetic modifiers of cancer risk (CIMBA)

The Consortium of Investigators of Modifiers of *BRCA1* and *BRCA2* (CIMBA) is an international collaboration initiative set up to investigate genetic modifiers. This consortium includes several consortia with similar aims, for example EMBRACE (Epidemiological study of *BRCA1* and *BRCA2* mutation carriers) and Modifiers and Genetics in Cancer (MAGIC). The creation of CIMBA allows for studies to be conducted with improved accuracy as collaboration results in increased sample numbers in studies. CIMBA includes groups that have access to a minimum of 100 *BRCA1* and *BRCA2* positive women (with or without cancer) and are able to provide data on genotype, phenotype and epidemiological risk in those samples. Groups follow standardised protocols for SNP genotyping.

Genome wide association studies (GWAS) have identified SNPs in the general population that affects risk of developing ovarian cancer. One SNP rs3814113 at 9p22.2 is linked to a decreased risk in ovarian cancer (Goode et al 2010). Within *BRCA1* and *BRCA2* mutation carriers this SNP also lowers the risk of ovarian cancer. There are a number of signs that indicate the existence of genetic modifiers to ovarian cancer risk. These include variability seen in penetrance figures both within mutation positive pedigrees and between different families (Milne & Antoniou 2011). Milne & Antoniou (2011) review the evidence on genetic modifiers in *BRCA1* and *BRCA2* positive breast and ovarian cancer. For example, they report that 2 SNPs (rs8170 and rs2363956) located at 19p13 are linked to increased risk of ovarian cancer or breast cancer within the general population and an increased risk of breast cancer only in patients positive for *BRCA1*. Interestingly, when breast and ovarian cancer are examined together these SNPs do not affect risk. Much of the work of CIMBA group members is currently being conducted and more evidence is required to confirm that 19p13 is linked to increased risk of ovarian cancer in women positive for mutations in *BRCA1*.

A further candidate for a genetic modifier is the polymorphism known as CASP8-D302H, which appears to result in a reduced risk of ovarian cancer in patients with *BRCA1* mutations. This reduction is estimated to be around 30%. This effect is only seen in patients with *BRCA1* mutations not in *BRCA2* (Engel et al 2010)

1.4.7 Hormonal and environmental modifiers of ovarian cancer risk in *BRCA1* and *BRCA2* women

Barnes & Antoniou (2011) report on a number of studies that examine non-genetic modifiers. They find that several factors can alter the cancer risks. These factors include oral contraceptive use, which is shown to both increase and decrease risk of breast cancer in different studies. In ovarian cancer a decrease in risk is noted in women using oral contraceptives. Nulliparous women show an increase in risk of breast cancer. Risk-reducing salpingo-oophorectomy (RRSO) is demonstrated to markedly reduce the risk of developing breast cancer in women positive for mutations in *BRCA1* or *BRCA2*. Radiation exposure during chest x-rays is demonstrated to increase the risk of developing breast cancer in *BRCA1/2* positive patients. The risk of breast cancer in women with high mammographic density and *BRCA1/2* positive is estimated to be doubled compared to women with low mammographic density.

1.5 Molecular pathogenesis of epithelial ovarian cancer

1.5.1 Clinical features of epithelial ovarian Cancer – histological subtypes

There are a number of different histological subtypes of epithelial ovarian cancer: endometrioid, mucinous, clear cell, transitional and serous. These histological types represent very different diseases. However, the main bulk of ovarian carcinomas fall into the high-grade serous category. The site of origin of ovarian cancer appears to be mesothelium of the ovary that forms inclusion cysts within the surrounding stroma. It is these inclusion cysts that transform into malignant cells. The ovarian carcinoma first metastasises to surrounding structures including the pelvis, abdomen and later more distant regions. If ovarian carcinoma is detected early before metastasising outside of the ovary then survival rates are much improved. Treatments generally include surgery and chemotherapy to prevent this spread. One of the difficulties in treatment success has been the vast heterogeneity of the disease.

1.5.2 A model of histological categories of ovarian cancer

Kurman & Shih (2010) suggest a model that categorises ovarian cancer based on molecular genetic and morphological characteristics. This model suggests there are two types of ovarian cancer, type I and type II.

Type I tumours are slow developing and present at an earlier stage. They are often borderline tumours with characteristics of both benign cysts and carcinoma. In this category are the low-grade serous and low-grade endometrioid, clear cell and mucinous carcinomas. In terms of the genetic characteristics, most low-grade serous tumours have mutations in vi-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (*KRAS*), v-raf murine sarcoma viral oncogene homolog B1 (*BRAF*) and v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian) (*ERBB2*). Mutations in *TP53* are very uncommon in type I tumours. In the low-grade endometrioid tumours mutations in catenin (cadherin-associated protein), beta 1 (*CTNNB1*), phosphatase and tensin homolog (*PTEN*), and phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA*) occur commonly, whereas mucinous tumours show mutations in *KRAS*.

Type II tumours are high-grade serous and high-grade endometrioid and also tend to be undifferentiated. They usually present at a later stage and are aggressive tumours. High-grade serous tumours have mutations in *TP53* in the majority of cases. Interestingly, type II tumours only very rarely show mutations in those genes commonly found occurring in type I tumours.

1.5.3 The cell of origin of epithelial ovarian cancer

Debate on the origin of epithelial ovarian cancer (EOC) continues. Discerning the cell of origin of EOC is relevant in research as this may shed further light upon pathogenesis of the disease and assist in earlier detection and development of new biomarkers. To date no clear precursors to the disease are discovered. Whilst the site of origin of ovarian cancer appears to be the mesothelium of the ovary, the cell of origin is still debated. The starting point is that the cell of origin is also the mesothelial cell of the ovary and that metaplastic alterations result in the transformation of these cells into the different histological subtypes of endometrioid, serous, mucinous, transitional and clear cell. The endometrioid subtype has the appearance of endometrial epithelium and the serous subtype is similar to the epithelium of the fallopian tube. Clear cell tumours

look like the epithelium of the gastrointestinal tract, with mucinous like the epithelium of the endocervix.

The cervix, endometrium and fallopian tubes originate from the Mullerian ducts, whereas the ovaries arise from the urogenital ridge, which is composed of mesothelium. Thus, it can be suggested that tumours resembling tissues derived from Mullerian ducts are not derived from the surface epithelium of the ovary, but from the columnar epithelium of the Mullerian tissues. It is possible then, that high-grade serous carcinomas originate from the fallopian tube and metastasise locally first to the ovary. Kurman & Shih (2010) review theories on the ovarian cancer cell of origin in an attempt to further elucidate the pathogenesis of ovarian cancer.

1.5.4 Mesothelium as the cell of origin

Histologically, the subtypes, serous, mucinous, endometrioid and clear cell do not look as though they have originated from the mesothelium of the ovary. One explanation for this is that metaplastic alterations of inclusion cysts arise following the invagination of the mesothelium of the ovary within the ovarian stromal tissues. These metaplastic changes may result in the transformation of the mesothelial cells to Mullerian duct type epithelial cells (Kurman & Shih 2010).

1.5.5 Mullerian duct as the origin of ovarian cancer

Precursor lesions in paratubal and paraovarian cysts that appear like serous, endometrioid, and mucinous or clear cell have not been discovered. In addition, mucinous type cancers appear to be more like gastrointestinal cells than Mullerian.

A number of studies, in women with a genetic susceptibility (i.e. *BRCA* mutation positive) to ovarian cancer demonstrate that in fact early lesions appear in the fallopian tube in this group of women. Later it was discovered that the majority of sporadic high-grade serous tumours display some involvement of the tubal mucosa. Thus, it is possible that serous tubal intraepithelial carcinoma (STIC), which is thought to derive from the fimbria, may also be the origin of high-grade serous carcinoma. In addition, the vast majority of STICs reveal mutations in *TP53*.

Louis Dubeau (2008) has studied the cell of origin for ovarian cancer extensively. He purports that epithelial ovarian tumours, rather than arising from the ovarian

mesothelium, are in fact derived from the tissues of the Mullerian duct. He goes further to suggest that the primary tumours of the fallopian tube, peritoneum and ovary are, essentially one single disease with Mullerian origins. Dubeau (2008) sites a number of arguments to support his theory including molecular biology, embryology and tumour morphology. Dubeau (2008) proposes that these tumours should be reclassified as 'extrauterine Mullerian cystadenomas or carcinomas' and that these should be sub-classified based on histology (i.e. serous, mucinous, endometrioid, clear cell). In terms of those patients with a mutated *BRCA1* or *BRCA2* gene this re-classification could have implications in prophylactic treatment in which cases surgery could exclude some of the ovarian cortex thus enabling women to remain fertile.

1.5.6 Gene Variants and Tissue Types

EOC is highly heterogeneous. Kurman & Shih's (2010) model, based on two main types of ovarian cancer, may well be oversimplified. EOC is likely to be several different diseases that share a similar anatomical location. The molecular genetics of type I and type II tumours are distinctly different. Type I tumours rarely display mutations in *TP53* and >60% of low-grade endometrioid tumours exhibit mutations in *KRAS*, *BRAF* and *ERBB2*. Alterations of the gene encoding β-catenin *CTNNB1* or *PTEN* have also been found in a number of studies (Bell 2005). By contrast, type II tumours such as high-grade serous carcinoma (HGSOC) commonly display mutations in *TP53* and amplification of the gene cyclin E1 (*CCNE1)*. The Cancer Genome Atlas (TCGA) data revealed that within type II tumours the different subtypes also display distinct molecular signatures. TCGA researchers conduct exome sequencing on the tumour DNA of 316 high-grade serous ovarian cancer (HGSOC) specimens comparing the tissue samples of matched normal control samples. They find that *TP53* is somatically altered in >96% of samples and that 9% of cases exhibit a germline mutation in *BRCA1* or 8% in *BRCA2*. In addition somatic alteration in *BRCA1/2* was noted in another 3% of samples. These data show that in ~50% of EOC, genes in homologous recombination (HR) are somatically altered.

Type I tumours rarely show mutations in genes mutated in type II tumours and vice versa. Mutations in *KRAS* are detected in both borderline and mucinous tumours; however, *TP53* mutations are very rare in these tumour types. Further elucidation of the molecular signatures of EOC are likely to lead to more accurate and detailed models of the disease and lead to more accurate, effective and targeted treatments for the disease.

1.5.7 The effect of inherited *BRCA1* and *BRCA2* mutations on pathology of EOC

Lakhani et al (2004) examine the histopathology of tumour samples of patients with *BRCA1* or *BRCA2* mutations as many studies have previously reported conflicting results. They investigate two groups of tumours: one includes 223 tumours from women with family history and the other 235 tumours from women unselected for a familial link. The family history cases have a minimum of one first-degree relative (FDR) or one second-degree relative (SDR) with ovarian cancer. Within the family history group 173 women are positive for mutations in *BRCA1* and 29 have a mutation in *BRCA2*, the remaining number (16) are not positive for a mutation in *BRCA1* or *BRCA2* and are excluded for the rest of the study. The mutation positive carriers include only those that have a deleterious mutation as defined by the Breast Cancer Information Core (BIC). These deleterious mutations result in a truncated protein product due to a frameshift or nonsense mutation. Also, included are large rearrangements, splice site variants or missense mutations that are clinically relevant as assessed by the BIC.

Tumour samples are assessed for histological subtype and tumour grade as well as the evidence of psammoma bodies (clusters of calcium seen through a microscope), vascular involvement, necrosis, mitotic index and amount of solid tumour. In addition, tumour samples are examined immunohistochemically for p53 and *ERBB2* (also known as *HER2*). In terms of age the patients with *BRCA1* mutations 84% are between 30-59 years old with only 16% above 60 years old. For *BRCA2* mutation carriers, 48% are 40-59 years old with the remaining 52% being over 60 years old. Tumour grade and histological subtype varies between *BRCA1* and *BRCA2* mutations carriers and controls. Samples from patients with *BRCA1* mutations are more often found to exhibit tumours of serous histology (44% of tumours) compared to controls (31% had serous histology). This translates to an odds ratio (OR) of 1.84 in patients with *BRCA1* mutations having a serous histological result. *BRCA2* positive patients have the highest level of serous histology tumours at 48%. Tumour grade also varies between groups. 72% of *BRCA1* mutation positive patient samples exhibit tumours that are high grade (poorly differentiated or undifferentiated) compared to 81% of *BRCA2* mutation positive patient samples and 55% of controls. It is noted that the *BRCA2* mutation positive patients include a group of women who are substantially older than the *BRCA1* group and this could have an affect upon tumour grade. They find *TP53* immunohistochemical staining is more often strong in *BRCA1/2* tumours and this validates similar findings by Ramus et al (1999).

Evans et al (2008) perform a large study examining ovarian cancer within families. They find that pathology of tumours affects the rate of identification of gene mutations. They combine the results of five studies and find that mucinous tumours very rarely exhibit mutations in *BRCA1* or *BRCA2*; the combined total for these studies is only 2% mucinous and 2% for borderline.

Piek et al (2001) conduct research to review the histopathology of fallopian tubes that are surgically removed from women that had a genetic susceptibility to ovarian cancer. They find that fallopian tubes demonstrate dysplastic and hyperplastic lesions and alterations in proteins involved in cell-cycle control and apoptosis are often evident. This implies a possible phenotype of a pre-malignant stage.

1.5.8 Effect of inherited mutations in *RAD51C* or *RAD51D* on pathology of EOC

How inherited mutations *RAD51C* or *RAD51D* affect the pathology of ovarian is yet to be elucidated. There appears to be no existing research on specific histology of tumours of women with germline mutations in either of these genes. It could be postulated that these might be more likely to be high-grade serous adenocarcinomas as they are known to interact with the *BRCA1* DNA repair network and as they are involved in DNA repair via homologous recombination. This research may go some way in revealing insight into this area as cases in these studies are enriched for high-grade serous adenocarcinoma.

1.6 Clinical relevance

1.6.1 The effect of gene variants on survival and chemosensitivity in ovarian cancer

The assessment of clinical relevance of variants in *BRCA1/2* in ovarian cancer cases reveals inconsistent data. In an effort to resolve this researchers (Yang et al 2011) use data from The Cancer Genome Atlas (TCGA) to examine the effect of harbouring faulty *BRCA1* or *BRCA2* genes on overall survival, progression free survival and chemosensitivity. Patients are considered to have a faulty *BRCA1* or *BRCA2* gene if found to be positive for a deleterious mutation or with hypermethlyation of the promoter region of either gene. Overall survival is defined as the period of time from the original surgery to resect the tumour up to death and progression free survival is defined as the period of time from original surgery up to recurrence or progression of disease.

1.6.2 Survival analysis in *BRCA1* and *BRCA2* mutation carriers

BRCA1 or *BRCA2* mutation positive patients with high-grade serous ovarian carcinoma show improved survival to those negative for mutations in either gene; those with mutations in *BRCA2* show a better survival rate than those with *BRCA1* (Yang et al 2011, Bolton et al 2012). Interestingly, those patients with epigenetic alterations in *BRCA1* that result in gene silencing show survival rates equivalent to those negative for a mutation in either gene.

1.6.3 Survival in novel variants

As yet, there is very little survival data on new variants (for example *RAD51C* and *RAD51D*) that are found related to ovarian cancer. Walsh et al (2011) do not find a significant association between overall survival and mutation status amongst their cohort. However, they do find a general tendency to better survival in those with mutations than those without. As more studies are conducted and statistics combined for newly discovered variants, survival in novel variants will be more accurately estimated.

1.6.4 Chemosensitivity in patients with *BRCA1* or *BRCA2* mutations

Yang et al (2011) investigate chemosensitivity by examining at two main points: 1, primary response to treatment (platinum based chemotherapy) and 2, time interval free of treatment after initial primary response. In the primary response a patient can be described as either chemo-sensitive or chemo-resistant depending on whether she has a response (partial or complete) to chemotherapy or no response (i.e. disease remained or progressed). Yang et al (2011) find that in High Grade Serous Ovarian Cancer (HGSOC) patients are more chemo-sensitive if positive for *BRCA2* mutations compared to patients either *BRCA1* positive or *BRCA* negative.

1.6.5 Targeted chemotherapeutic treatments

Detection of novel gene variants may lead to a greater understanding of the biological pathways involved and result in the identification of new treatment targets. This has already been the case for patients with *BRCA1* or *BRCA2* related breast cancer. In addition, many *BRCA1* breast cancers are triple negative. Interestingly, TCGA notes the parallels between basal-like breast cancers (that are generally triple negative) and

high-grade serous ovarian carcinoma. Uncovering the gene variants specific to this subtype might reveal that the chemotherapeutic treatments are also transferable.

1.6.6 PARP inhibitors

Poly (ADP-ribose) polymerase (PARP) is a DNA repair enzyme with a specific role in ssDNA break (SSB) repair. It initiates a signaling cascade on detection of SSBs. Chemical inhibition of one of the two isoforms (PARP1 and PARP2) has been shown to be lethal only to those cells that are deficient in proteins of DNA repair via homologous recombination (for example, *BRCA1* and *BRCA2*). The action of PARP inhibition (Figure 1.5, page 46) works simply by forcing homologous recombination deficient cells to go through an alternative DNA repair pathway; cells without this genetic deficiency follow the normal repair pathway and are spared (Ashworth 2008). The HR deficient tumour cells only are targeted and result in the selective death of those tumour cells. In addition, PARP inhibitors, currently in phase III clinical trials, show efficacy on tumours that inherently express *BRCA1/2* deficient characteristics. This phenomenon is known as BRACness (TCGA). Those tumours with epigenetic silencing of *BRCA1* are also sensitive to PARP inhibitors.

Figure 1.5. *BRCA1* or *BRCA2* deficient cells are sensitive to PARP inhibitors; normal cells are not

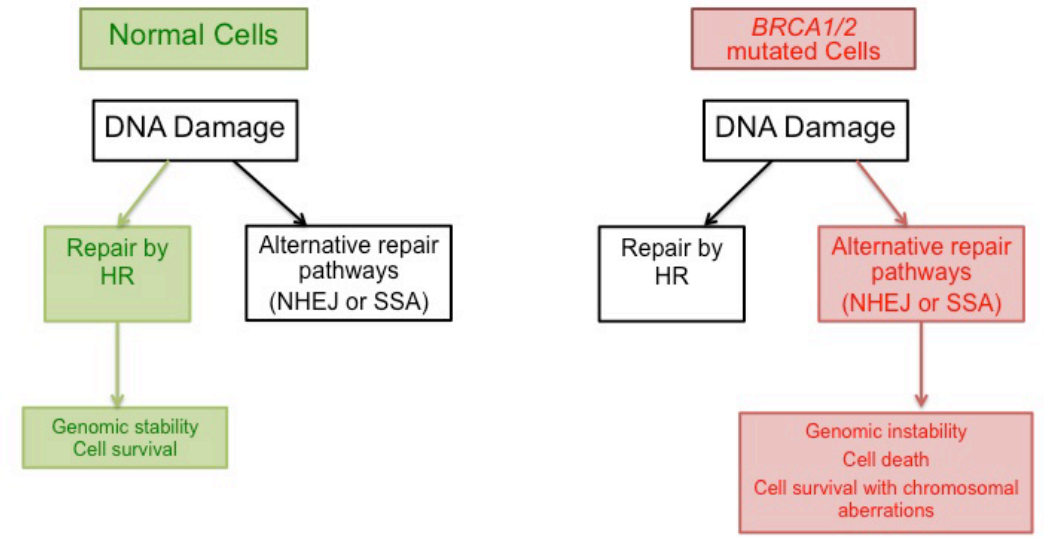


***Figure 1.5. PARP inhibitors**. PARP inhibitors affect BRCA1 or BRCA2 deficient cells only, because following DNA damage cells are forced into alternative DNA repair pathways. These pathways are error prone and result in either the death of BRCA deficient cells or in their survival with gross chromosomal aberrations. Normal cells are not sensitive to PARP inhibitors as they are able to follow the homologous recombination pathway following DNA damage (Ashworth, A 2008 J Clin Oncol 26:3785-3790). HR=homologous recombination, NHEJ=non-homologous end joining, SSA=single-strand annealing pathway.*

A research group has designed a laboratory test that can assay the sensitivity of cells to PARP inhibition. Mukhopadhyay et al (2010) develop a laboratory test that can predict the sensitivity of cells to PARP inhibition. The assay is developed to test EOC tumour cells for deficiencies in Rad51 foci formation. They find that 93% of cells deficient in homologous recombination are sensitive to PARP inhibition and that failure to form Rad51 foci is a good predication of homologous recombination deficiency.

1.7 An overview of DNA sequencing technology for mutation detection

Frederick Sanger won the Nobel Prize in Chemistry in 1980 for developing one of the first methods in DNA sequencing; the 'Sanger Method' is also known as the chain termination or dideoxy method and came into use in 1977. DNA sequencing involves copying the original DNA sequence numerous times. The original Sanger method is a four-tube reaction; into each tube DNA template, primer, polymerase, free nucleotides and one of 4 radioactive dideoxynucleotides are added. First the template DNA is denatured into single stranded DNA. The primer is hybridised to the template strand and is the starting point for sequencing. The DNA polymerase anneals to the primer and synthesises another DNA strand complimentary to the original template by adding the free nucleotides in the tube. Since the tube also contains one of the radioactive dideoxynucleotides the growing DNA chain is terminated as this modified nucleotide is missing the 3'OH group required to form the phosphodiester bond with the next nucleotide. This results in numerous copied fragments of DNA of different lengths that end in a radioactive dideoxynucleotide (ddNTP). Each tube has a different ddNTP; thus when the reaction products of each tube are separated on adjacent lanes on an electrophoretic polyacrylamide gel the length of fragment (smaller DNA molecules migrate faster) indicates the position of the nucleotide and the fluorescence indicates the base. The gel is visualised under ultra-violet light or by autoradiography and read from the bottom upwards (Sanger et al 1977).

Determining the order of the four bases in DNA heralds a new era in which knowledge of genetic variation is now more accessible, opening up greater opportunity to discern disease aetiology and epidemiology; and drive new insights into human evolution. DNA sequencing methodology has evolved rapidly over the last decade, during which time the draft reference sequence of the human genome was sequenced by the huge international collaborative efforts of the Human Genome Project in 2001. Figure 1.6 depicts the progress of DNA sequencing technology throughout the last 10 years.

Figure 1.6. The progress of DNA sequencing during the last 10 years.

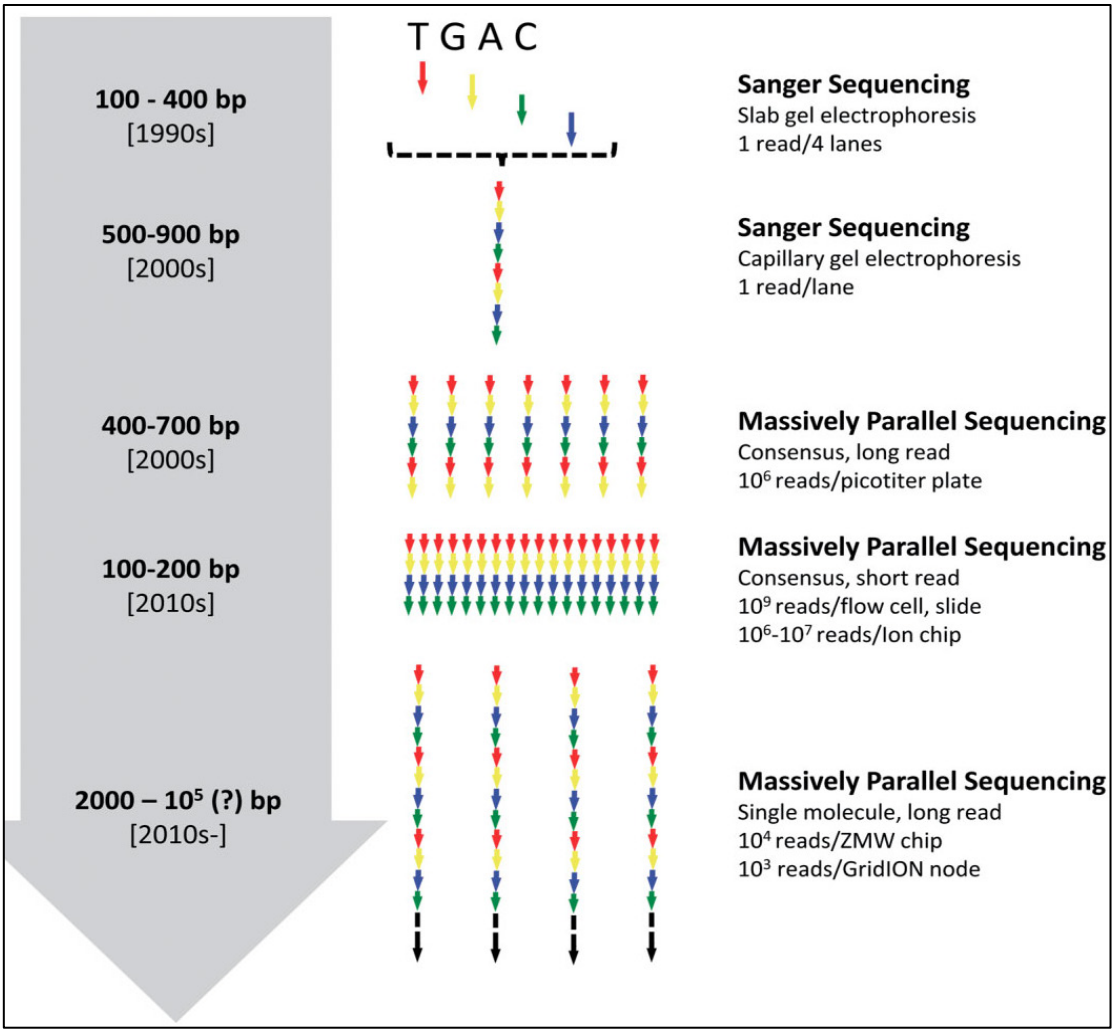


***Figure 1.6. The progress of DNA sequencing during the last 10 years**. This image is a graphical representation of the rapid progress made in DNA sequencing technology throughout the last decade. Adapted from: Strannenheim & Lundeberg (2012) Stepping stones in DNA sequencing. Biotechnol J. 2012 Sep;7(9):1063-73*

1.7.2 DNA sequencing by capillary electrophoresis

DNA sequencing by Capillary Electrophoresis, still recognised as the gold standard, is a modification of the Sanger Method. It employs dye-terminating chemistry to allow for one sequencing reaction to take place instead of four. The fragments of copied DNA are separated by capillary electrophoresis and each of the 4 ddNTPs is fluorescently labelled with a different coloured dye; laser excitation enables their identification. The trace output produced is a four-colour system that represents each of the 4 bases (Figure 1.7 overleaf).

Figure 1.7 Capillary electrophoresis trace output.

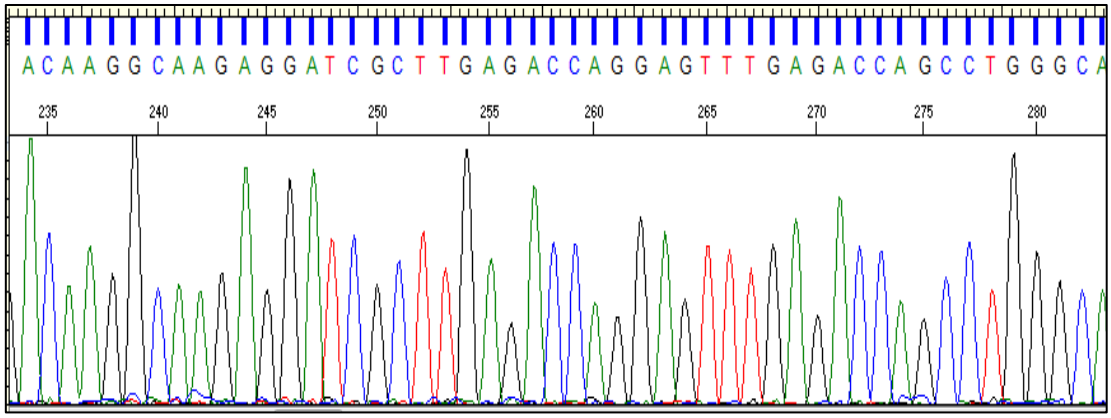


Figure 1.7 Capillary electrophoresis trace output. *This image shows the output from DNA sequencing using capillary electrophoresis. Red = Thymine, Blue = Cystosine, Black = Guanine and Green = Adenine.*

1.7.3 Next generation sequencing (NGS)

Next generation sequencing, also known as massively parallel sequencing or second-generation sequencing, can effectively sequence DNA templates in the order of kilobases (Kb) to megabases (Mb) simultaneously. Up to 600 Gb per run in around 10 days are achievable with the latest Illumina HiSeq2000; this, coupled with a high-throughput per run, results in a highly cost effective sequencing method (Mamanova et al 2010). Today's NGS technology is able to sequence more than 5 whole human genomes simultaneously at a depth of 30 X with the latest dual flow cell Illumina HiSeq2000 or sequence 100 exomes in a single run. However, when sequencing large numbers of subject DNA samples and if the researcher's interest is in only a few genes this could be highly wasteful. To circumvent this problem target enrichment approaches are employed to isolate and enrich the genomic region of interest and enable just those regions to be sequenced. This increases sample throughput to unprecedented scales, depending on the size of genomic region targeted.

1.7.4 Next generation sequencing system technologies

These systems vary in terms of sequencing chemistry, throughput and cost. All have similar library preparation protocols, during which the template sample is prepared for sequencing.

1.7.5 Illumina Genome Analyzer IIx (GAIIx) and Illumina HiSeq2000

Figure 1.8. The Illumina HiSeq2000

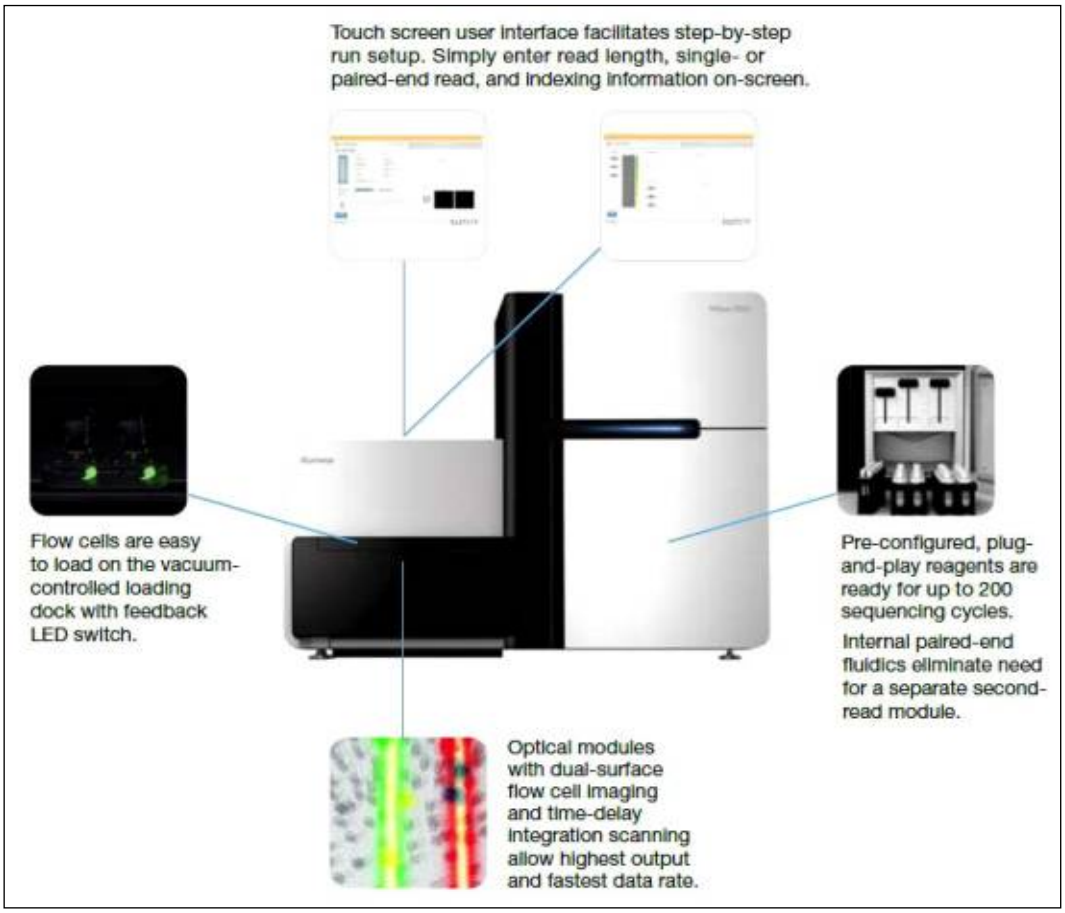


Figure 1.8. The Illumina HiSeq2000. *This image describes some of the features in this technology including the dual flow cell and streamlined plug and play reagents. Nevertheless, this instrument requires a large amount of space in the laboratory and a separate C-bot instrument for cluster generation. Image sourced from http://www.illumina.com/systems/hiseq_2000_1000.ilmn.*

The Illumina HiSeq2000 system is the successor to the GAIIx. This instrument amplifies by generating clusters of the original template immobilised on a glass slide (known as a flow cell). The clusters are sequenced using fluorescently labelled reversible dNTPs and 'sequencing-by-synthesis' to produce 150bp reads at 600 Gb data per run at a rate of 81 Gb per day. This new sequencer has a dual flow cell system to allow for reading sequences from both sides of the flow cell. Essentially, this doubles the number of reads (a detected string of nucleotides) to enable sequencing of more genes and more samples. The HiSeq2000 also has two independent flow cells, which allows for the simultaneous running of different experiments. Thus, one cancer genome and one whole genome can be sequenced in just one week. The new Illumina system also improves throughput. More streamlined sample preparation involves master mixed reagents, the elimination of 75% of the purification steps and simplified

indexing protocols. At the end of 2010 new robotics systems were introduced for library preparation, which enables 400 samples to be prepared in just one week. These changes to the system are improving scalability of the method as well as accuracy due to the minimisation of pipetting errors.

1.7.6 Roche 454 system

Produces on average 700bp reads at 700 Mb data in a 23 hour run. It uses capture bead-based technology, on which one fragment of DNA template is immobilised. These beads are amplified by emulsion PCR, which results in several million copies of each fragment per bead. The beads are then added to a PicoTiter plate that will only accept one bead in each well. Then sequencing enzymes are added and the nucleotides are washed over in a set sequence and if one is complimentary to those on the beads, the Charge Couple Device (CCD) camera detects fluorescence. Data analysis is performed examining the intensity of signal and the position of the signal on the PicoTiter plate (Tucker et al 2009).

1.7.7 SOLiD (sequencing by oligonucleotide detection)

This system by Applied Biosystems produces 300 Gb of data per run in 14 days for paired end (sequencing each end of a DNA sequence in both forward and reverse directions) 75bp reads. SOLiD uses bead based technology and emulsion PCR. Sequencing is achieved by ligation of 4 differently labelled ligation probes that ligate to the sequencing primer competitively; the sequence is determined by identification of the first and second base in each reaction. A ligation-detection-cleavage cycle is repeated numerous times to extend the sequencing primer; these multiple cycles are repeated again with a universal primer following cleavage of the first primer. This results in each base being detected twice, thus it has an inbuilt error correction system (Tucker et al 2009).

Massively parallel sequencing can also be achieved by sequencing single DNA molecules in a system known as Helicos. This system does not amplify the original template, but instead sequences each molecule and identifies it via fluorescence immediately.

1.7.8 Complete Genomics

The Complete Genomics method sequences whole genomes by technology known as sequencing by unchained ligation. Template DNA is initially fragmented into 500bp segments into which four adapters are included using intramolecular ligation and restriction enzymes that cut at repeat intervals. This produces 'DNA nanoballs', which are amplified circles of ssDNA that are exact copies of the original template. This method of amplification is known as restriction/circularisation or rolling circle amplification (RCA). The DNA nanoballs are hybridised to the surface of a DNA nanoball microarray, which contains complimentary DNA. The DNA sequence is determined by fluorescence as the complementary DNA contains fluorescent detection probes that are hybridised to the anchor probes. The anchor probes search for the four adapters within the DNA nanoballs (Strannenheim & Lundeberg 2012).

1.7.9 Ion Torrent

The Ion Torrent system works differently to the previous methods as it uses an electrical detection system as opposed to fluorescence. The Ion Torrent library preparation is similar to the Roche 454 system in which the template is amplified using a bead based emulsion PCR method. The beads from the emulsion PCR are added to the wells of an ion chip, which has the ability to distinguish free protons. Hydrogen ions are released when DNA polymerase adds nucleotides, this leads to a change in the free protons, which is distinguished by a sensor in the ion chip and converted to an electrical signal (Strannenheim & Lundeberg 2012).

Each of these systems has advantages and disadvantages. For example, Illumina and SOLiD produce many short reads and this has clear disadvantages compared to the longer read length of the 454 system; in that a proportion (around 10-20%) of the reads produced by Illumina are not of sufficient quality to be usable. SOLiD has a similar problem with a fraction of reads not reaching usable quality levels. However, this issue can be overcome by increasing read depth. The cluster amplification of template strands on the Illumina system produces 1,000 copies of the original template, enabling the simultaneous sequencing of the same DNA strand. The 454 system produces fewer longer reads of higher quality ~95% align to the reference sequence (Harismendy et al 2009). This system is however, a lot more expensive per base, more than 10 times more expensive in fact (Tucker et al 2009). A major disadvantage of the Complete Genomics system is that it is not available for sale; the company

keeps the system as an in-house platform offering DNA sequencing services. Thus, the system is not a financially competitive one. One obvious advantage of Ion Torrent is the use of an electrical detection system in comparison to detection by fluorescence as this can massively reduce costs by eliminating the requirement of optics and modified nucleotides. In addition, data stored as fluorescent images requires additional processing to become meaningful. One clear disadvantage of this system is that it produces very short reads (200 bp) and errors in detecting indels can be evident (Strannenheim & Lundeberg 2012).

1.7.10 Third generation sequencing systems

A new generation of DNA sequencing threatens to supersede massively parallel sequencing. These technologies use single molecule DNA sequencing, which has the advantage of longer read lengths (> several hundred base pairs) enabling easier read alignment and sequence assembly and improved accuracy in variant detection. This coupled with greatly reduced run times; result in even higher likelihood for the use of this technology for genetic screening. Two main companies are developing third generation technologies. These are PacBio and Life Technologies, each of which has their own advantages and disadvantages. Both technologies utilise the same camera technology to record data output. The charge coupled diode (CCD) system uses CCD array technology that currently has a limited size and results in throughput no higher than that of Illumina or SOLiD technologies (Munroe & Harris 2010). At the present time, most laboratories do not have this technology available.

1.7.11 Challenges for second-generation sequencing technology

The last decade has seen massive progression in sequencing technology in terms of increased throughput, reduced cost and time. However, this progress brings additional challenges in that the library preparation and bioinformatics required needs to keep up with this accelerated advancement in technology. This has produced bottlenecks at each end of the sequencing process. Many of the library preparation issues are now being addressed, with increased multiplexing, sample barcoding and library preparation automation systems. The downstream analyses, bioinformatics and data storage are still demanding, however, improvements are being introduced in these areas. Certainly, the most important first step in sequence analysis is the mapping stage (also known as read alignment to the reference genome) and this requires

greater speed and accuracy than at present to be able to use the full capacity of current sequencing technologies.

1.7.12 Whole exome sequencing

Whole exome sequencing is the sequencing of the coding sequence of all protein coding genes. Jones et al (2009) identify *PALB2* as a susceptibility gene in pancreatic cancer using whole exome sequencing. They narrow down the search to three candidates as these three contain deleterious mutations with the loss of both alleles. The likely causative mutation is in partner and localiser of *BRCA2* (*PALB2*) gene, as the other two serpin peptidase inhibitor, clade B (ovalbumin), member 2 (*SERPINB2*) and renal tumour antigen (*RAGE*) are found to typically contain stop codons within the general population and *PALB2* is already linked with breast cancer susceptibility. In addition, *PALB2* is one of the Fanconi Anaemia pathway genes (also known as *FANCN*). Once the likely candidate is identified, Jones et al (2009) use DNA sequencing to examine *PALB2* in a larger group of pancreatic patients whom are enriched for family history. Jones et al (2009) conclude that *PALB2* is probably a pancreatic cancer susceptibility gene as it is commonly mutated in hereditary forms of the disease. Thus, whole exome sequencing is one approach to the identification of cancer predisposition genes. Exome sequencing in cancer cases and controls may be an appropriate method for the detection of new cancer predisposition genes, however, in the clinical setting it may be more economically and practically feasible to sequence the specific genes that confer high disease risk for the purposes of risk prediction and early detection of disease.

1.8 Genetic testing for familial ovarian cancer

1.8.1 UK Guidelines for genetic testing

NICE (National Institute for Health and Clinical Excellence) is an organisation set up to independently issue guidelines and quality standards for NHS practices. The first familial breast cancer guidance was delivered in 2004, known as NICE clinical guidance [CG14], and this focuses on the care of women with an increased risk of breast cancer. This first guidance defines care at primary, secondary and tertiary levels of the health system. This was further updated in 2006 when guidance was issued on clinical screening management of women with an elevated breast cancer risk.

This guidance is concerned with providing care for women with an elevated risk of developing breast cancer due to family history of the disease or other relevant cancers (sarcoma in family member under 45 years, glioma or childhood adrenal cortical carcinomas and families with multiple cancers at an early age of onset).

The key points in the NICE 2006 [CG14] guidance concerning policy for clinical management of women with elevated breast cancer risk are summarised in Table 1.1.

Table 1.1 Key policy points for clinical management of women with elevated breast cancer risk

Assessed level of risk	High	Raised/moderate	Population
Appropriate level of care	Referral to a specialist genetic clinic in tertiary care. Genetic counselling offered.	Secondary Care	Primary Care
Clinical management	Women offered annual MRI 30-39 years	Women 40-49 offered annual mammography and/or annual MRI	None

Table 1.1 Key policy points for clinical management of women with elevated breast cancer risk*. NICE clinical guidance CG41.*

An assessed high-risk level includes women with a 10 year risk >8% at 40-49 years and an overall lifetime risk of >30%; this risk level also includes women with a >20% probability of being a mutation carrier in one of the genes *BRCA1*, *BRCA2* and/or *TP53*. A moderate risk level includes those women with a 10-year risk level of 3-8% at 40-49 years and an overall lifetime risk of 17-30%. Population risk level is considered to be a 10-year risk of <3% and an overall lifetime risk of <17%. It is not recommended that women with elevated risk should be sought, that only those women whom approach primary care with anxieties about their risk level should be assessed. In the first instance a family history is taken looking at both first and second-degree relatives (NICE 2006 CG14). The guidance was further updated in June 2013 [CG164]. Now genetic testing is extended to those without a cancer diagnosis or a living affected relative if they have an assessed 10% probability of a mutation in one of the high penetrant genes (*BRCA1*, *BRCA2* or *TP53*).

1.8.2 Intervention strategies and risk reduction

The strategies for clinical intervention in women assessed as high risk include regular screening, chemoprevention and risk reducing surgery. For breast cancer risk surgery

includes mastectomy or annual MRI. However, for ovarian cancer risk current adequate screening is not available so risk-reducing salpingo-oophorectomy (RRSO) is performed if women have completed their families and have a mutation in *BRCA1* or *BRCA2* (Domcheck et al 2010).

Domcheck et al (2010) perform a prospective analysis on a large cohort of 4,255 *BRCA1* and/or *BRCA2* positive women that received risk-reducing surgery. They examine mutation type, type of surgery (mastectomy or RRSO) and cancer history. The main findings of the study are summarised as follows:

1. In women who have risk-reducing mastectomy none have a breast cancer diagnosis in the subsequent 3 years of follow-up; this compares to 7% cancer diagnoses in women whom do not have mastectomy.

2. In women who have RRSO the hazard ratios (HR) for *BRCA1* mutation carriers are 0.31 and no *BRCA2* carriers have a cancer diagnosis during 6 years of follow-up; this compares to 3% cancer diagnoses in women whom do not have RRSO.

3. In women that have RRSO and are previously diagnosed with breast cancer the HR are 0.15 reduction in risk in *BRCA1* mutation positive patients; and zero further cancer diagnoses are confirmed in women with *BRCA2* mutations.

4. In women that have never been diagnosed with breast cancer the HR associated with RRSO in *BRCA1* mutation positive women are 0.63 and 0.36 in those with *BRCA2* mutations.

5. If women have RRSO <50 years old the risk reduction HR is 0.51 however, if RRSO is performed >50 years there is no observed reduction in risk.

6. In both *BRCA1* and *BRCA2* positive women no reduction in risk is observed in the subsequent diagnosis of a second primary breast cancer.

7. Mortality is reduced in women who have RRSO and no previous breast cancer diagnosis (HR 0.45) this compares to those who have previously had a breast cancer diagnosis (HR 0.30).

1.8.3 Risk prediction models

A number of clinical criteria could indicate the presence of *BRCA1* or *BRCA2* mutation. These include early onset disease (under 50 years at diagnosis for breast cancer and under 60 years for ovarian cancer), diagnosis of contralateral breast cancer or ovarian and breast cancer, male breast cancer, triple negative breast cancer diagnosed under 50 years and in a specific population group (e.g. Ashkenazi Jewish).

Several models designed for the prediction of germline *BRCA1* or *BRCA2* mutation status have been developed since the late 1990s. The main models in current use are BRCAPRO, BOADICEA, the Myriad Prevalence tables and Tyrer-Cuzick models. These models have been extensively assessed in terms of validity in predicting the presence of mutations based on family history of breast and/or ovarian cancer of the proband. BRCAPRO and BOADICEA are similar models and appear to perform equally well in terms of both sensitivity and specificity (Schneegans et al 2011).

BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) is a genetic risk model for mathematically estimating the probability of the presence of a mutation in *BRCA1* or *BRCA2* and in addition the probability of developing breast and or ovarian cancer. BOADICEA does not estimate the probability of detecting a mutation, however as this also involves the sensitivity of the mutation detection method. The BOADICEA risk model is useful for assessing further need for genetic counseling or genetic testing. This model takes a polygenic approach to risk prediction (Antoniou et al 2004).

BRCAPRO is a similar Bayesian based model that uses family history of breast and ovarian cancer; it also includes family history of male breast cancer and diagnoses of bilateral breast cancer within the family (BayesMendel Lab 2012)

1.9 Research aims

1. Almost half of inherited EOC risk is due to moderate to high penetrant gene variants other than *BRCA1* and *BRCA2*. The aim of this research is to examine the contribution of six candidate genes to EOC.

2. Discovering gene variants within the missing 46% will lead to improved risk prediction and early detection. Using the most up to date sequencing technology to

detect these variants will assist in meeting this aim. There is a translational element to this research. If gene variants can be detected swiftly and cheaply then this will result in a larger population of women being tested; thus enabling superior risk prediction and earlier detection of disease (or disease potential).

1.10 Hypotheses

1. The missing proportion of genetic predisposition to inherited/familial ovarian cancer is due to rare variants in moderate to high penetrance genes.

2. Next generation sequencing of large sample sets of cancer cases and healthy matched controls is an appropriate method to detect rare variants in moderate penetrance genes.

3. Recently discovered rare variants in HR genes are validated in these data.

4. Targeted re-sequencing using NGS technology is ready and suitable for use in the Molecular Genetics Diagnostic laboratory.

Including novel rare variants in moderate to high penetrance genes that confer a high risk of ovarian cancer will give a more accurate prediction of ovarian cancer risk. Genetic risk models for mathematically estimating the probability of developing ovarian cancer should include these newly discovered variants.

Chapter Two

Establishing new technology: The application of next generation sequencing for the detection of germline gene mutations

2.0 Introduction

2.1 *BRCA1* and *BRCA2* in ovarian cancer susceptibility

Ovarian cancer is the predominant cause of death from gynaecological malignancy and ranked fourth in the UK for cancer death amongst women. Globally, ovarian cancer is the cause of death of ~125,000 women annually (Cancer Research UK statistics, 2009). The detection of ovarian cancer at an advanced stage results in a poorer prognosis for patients. Only a small proportion of ovarian cancers are detected in the early stages when survival rates are best. One third of women diagnosed present with distant metastases and the five-year survival rates for these patients are only 15% (Cancer Research UK statistics, 2009). Inherited ovarian cancers constitute ~10% of all invasive ovarian carcinomas (Stratton et al 1998). Penetrance estimates in *BRCA1* and *BRCA2* genes vary widely amongst studies and this variability is likely to be due to the cohort studied (i.e. from breast-ovarian cancer families or studies unselected for family history). Gayther (2012) reports that the cumulative risk of epithelial ovarian cancer to age 70 is between 40% and 50% for *BRCA1* mutation positive women and between 20% and 30% for *BRCA2* mutation positive women. Antoniou et al (2003) estimate ovarian cancer risks due to *BRCA1* or *BRCA2* genes by combining data from 22 studies that included cases not selected for family history. They conclude that the average cumulative lifetime (by age 70) ovarian cancer risk in *BRCA1* positive women is 39% and for *BRCA2* is 11%.

New methods that can detect ovarian cancer at an earlier stage are vital to improve survival rate for the disease. Next generation sequencing (NGS) technology is an advancing method that could be employed to identify quickly, cheaply and accurately biologically relevant genetic variants enabling early detection of women at high risk of developing ovarian cancer. This proof of principle study aims to re-sequence *BRCA1* in 12 ovarian cancer patients with known positive *BRCA1* mutations. Long Range PCR (LR-PCR) is used as a target enrichment strategy to amplify *BRCA1* from 12 ovarian cancer patient samples. The study successfully multiplexes 11 samples in one lane of an Illumina flow cell to re-sequence the whole of *BRCA1* in these 11 samples. I am blinded to the known mutations in each sample and all mutations are accurately identified. The study concludes that multiplexed sequencing protocols can be

employed for the targeted re-sequencing of multiple patient samples in each lane of the Illumina flow cell and that scaling up of these methods will ensure that this technology is ready for use in the diagnostic clinic.

2.2 The research questions

1. Can Massively Parallel Sequencing (MPS) technology be applied for use as a clinical tool for the detection of germline gene mutations in *BRCA1*?

2. Can MPS be employed as a research tool to identify additional rare variants that confer a moderate increase in cancer risk?

2.2.1 The research questions in context - why is the research important?

As women with germline mutations in *BRCA1* or *BRCA2* have a high lifetime ovarian (Antoniou et al 2003). The introduction of better genetic screening for the identification of women with inherited mutations is required to improve survival rates for the disease. Early disease detection benefits survival as detection at stage I disease the 5 year survival rate is ~90% (Cancer Research UK statistics 2006). Improved survival is achieved by earlier intervention with, for example, prophylactic or risk reducing surgery or early monitoring in patients identified at high risk.

The National Institute for Health and Clinical Excellence (NICE) guidelines on genetic testing report that women should be tested for *BRCA1* or *BRCA2* mutations if they have a 20% or more assessed risk of having a mutation (NICE 2006 recommendations). The guidance states that *'Genetic testing is only appropriate for a small proportion of women who are from high-risk families'*. This risk is assessed on family history of disease and only those with a living affected relative are tested. NICE guidelines [CG164] were recently updated to include those without cancer or a living affected relative if they have an assessed risk of 10% of a *BRCA1* or *BRCA2* mutation. In a population-based study Soegaard et al (2008) find that age is as significant a predictor of the presence of *BRCA1* or *BRCA2* mutations as family history. In patients found to have a mutation in either gene more than half are diagnosed with ovarian cancer under 50 years old compared to <20% diagnosed with cancer that are found to have a negative mutation status. Antoniou et al (2003) find that age is a significant risk factor for development of breast or ovarian cancer in women positive for mutations in *BRCA1* or *BRCA2*; and that two additional factors modified the risk of breast or ovarian cancer. The location of the mutation is relevant in altering the risk of developing disease and the year of birth also has a significant effect in that the earlier that women

were born the lower their risk of developing cancer. King et al (2003) investigate the impact of *BRCA1* and *BRCA2* mutations on breast and ovarian cancer risk. Interestingly, they find that 50% of their 1,008 breast and ovarian cancer patients with detected *BRCA1* or *BRCA2* mutations are from 'low incidence families' that have no disease in female relatives on their maternal side. They report that this is because these are small families and that the mutations are inherited from their paternal side. Walsh et al (2010) report that women, without an apparent family history, positive for *BRCA1* or *BRCA2* mutations are equally at high risk of ovarian or breast cancer as those with a strong family history. Family history can only arbitrarily be assessed and subsequently, many women with *BRCA1* or *BRCA2* mutations are not offered genetic testing.

Genetic screening will improve disease management by refining the classification system for serous carcinoma. The further sub-classification of High Grade Serous Carcinoma (HGSC) according to *BRCA* mutation status could lead to earlier detection and thus improve the prognosis for these specific ovarian cancer patients (Press et al 2008). Routine use of genetic testing for all women diagnosed with breast cancer or ovarian cancer could improve treatment options. The introduction of PARP inhibitors, for the treatment of BRCA associated tumours is likely to introduce a more targeted therapy for these specific tumours that should effectively reduce toxicity and increase specificity (Farmer et al 2005).

Dr Ranjit Manchanda of the Institute for Women's Health, University College London (UCL) is leading a trial into genetic screening for *BRCA1* and *BRCA2* mutations in the Ashkenazi Jewish population. The study known as Genetic Cancer Prediction through population screening (GCaPPS) will assess the viability of screening an entire population in order to determine who in that population has a higher risk of developing cancers related to germline mutations in *BRCA1* or *BRCA2*. The trial will examine the effects of population-based screening and weigh up these against the existing protocol of family history based screening. The results of this trial will be very interesting and may influence how the prediction of genetic susceptibility to cancer is assessed in clinical practice for a wider population. Genetic screening for a wider population would require in an increase in counselling on the results of that screening.

To date an expedient *BRCA1*/*2* mutation-screening tool has not been developed that is rapid and cheap enough per patient to allow for testing patients without family history. Next generation sequencing (NGS) technology could be applied to the detection of

BRCA1 and *BRCA2* as well as other gene variants linked to the genetic predisposition to ovarian cancer.

2.3 *BRCA1* and *BRCA2* Genes: structure and function

Figure 2.1. The structure of *BRCA1 and BRCA2* genes

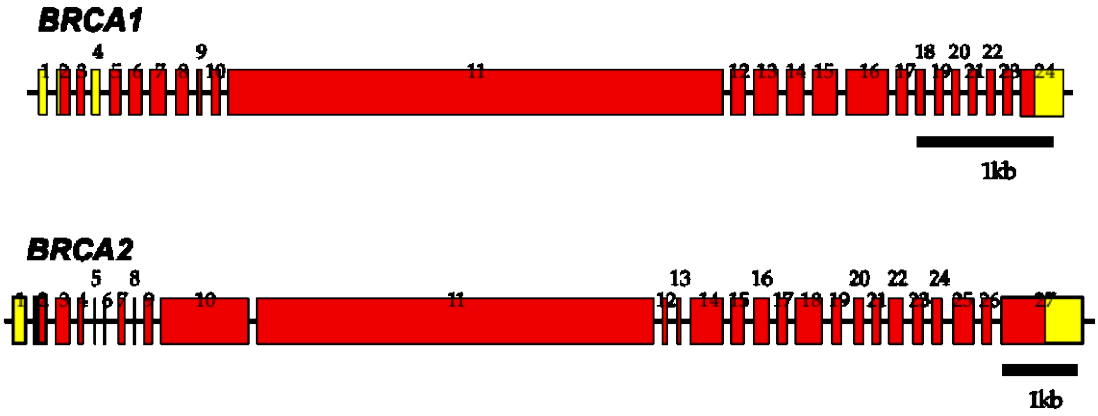


Figure 2.1 The structure of BRCA1 and BRCA2 genes. *This schematic representation shows the structure of BRCA1 and BRCA2 genes. Red blocks are coding exons and yellow represent non-coding exons*

BRCA1 is located at chromosome 17q21 and comprises a region 81 Kb in size with 24 exons, 22 of which are coding exons. *BRCA2* is located at 13q12-q13 and comprises a region 84 Kb in size with 27 exons, 26 of which are coding exons (Figure 9). *BRCA1* and *BRCA2* have translational start sites in exon 2.

Brown et al (1996) found a genomic region of chromosome 17 that includes *BRCA1* with a 30 kb duplication, which leads to two copies of exon 1 and exon 2 of *BRCA1*, two copies of exon 1 and 3 of *NBR1* (Neighbour of *BRCA1* Gene 1) and in addition two copies of an intergenic region 295 bp in size; Brown et al (1996) suggest these multiple exons are in fact duplicated pseudogenes. A large proportion (41%) of *BRCA1* contains *Alu* repeat sequences that are 69-231 bp in length (Smith et al 1996); and these occur approximately every 650 bp throughout the entire sequence of the gene (Tancredi et al 2004). The size of the introns of *BRCA1* varies between 403 bp to 9.2 Kb.

BRCA2, like *BRCA1* has a very large exon 11. *BRCA1* and *BRCA2* contain AT rich sequences in their coding regions and translational start sites are situated in exon 2 of each gene (Tavtigian et al 1996). A region of *BRCA2* has been designated by Gayther et al (1997) as the 'ovarian cancer cluster region' (OCCR) and this region is situated between nucleotides 3035 and 6629 spanning 3.3 Kb of exon 11. Gayther et al (1997) found that breast and ovarian cancer families had a higher ratio of ovarian cancer

cases to breast cancer cases if mutations resided in this portion of *BRCA2*. Interestingly, the RAD51 binding domain resides in this region of *BRCA2* (Thompson & Easton 2001).

The protein product encoded by *BRCA1* is relatively large consisting of 1,863 amino acid residues (Huen et al 2010). At its amino terminal the BRCA1 protein has a conserved RING domain and at its carboxyl terminal are tandem BRCT domains. RING domains are recognised as regions involved in the ubiquitination of proteins, a process involving the tagging of proteins for their degradation. By contrast the BRCT domains are regions implicated in the binding of phosphorylated proteins. These domains are regularly observed in proteins concerned with the DNA damage response.

The functions of *BRCA1* and *BRCA2* continue to be elucidated. *BRCA1* is probably the better known of the two, with roles in the mechanisms of DNA repair, cell-cycle regulation, chromatin remodelling, protein ubiquitination and transcriptional regulation (Narod & Foulkes 2004). The diagram (Figure 2.2) depicts the role of *BRCA1* in the DNA damage response. DNA damage is detected via sensors. These sensors are kinases, for example Ataxia Telangiectasia Mutated (ATM) that once activated phosphorylate checkpoint kinase 2 (*CHK2*), which in turn phosphorylates *BRCA1* and results in the cessation of cell division. *BRCA1* further phosphorylates downstream targets involved in cell cycle control, including p53 and Rb. The protein products of *BRCA1* interact and form complexes with other proteins in the BRCA network to bring about its various functions. For example, *BRCA1* forms a heterodimer with BRCA1 associated RING domain 1 (*BARD1*) to result in ubiquitin ligation of downstream targets. S-phase or G2 arrest are brought about via the complexing of *BRCA2* and *RAD51*, which in turn interacts with Fanconi Anaemia complementation group D2 (*FANCD2*) and its subsequent annealing to *BRCA1* (Narod & Foulkes 2004).

Homologous recombination and transcriptional regulation are controlled by BRCA1-associated surveillance complex, which includes genes: Bloom syndrome, RecQ helicase-like (*BLM*), *MSH2-MSH6* and *MRE11-RAD50-NBS1*). *BRCA1* also has some involvement in non-homologous end joining, in the regulation of chromatin remodelling and in the regulation of apoptosis (Narod & Foulkes 2004).

The predominant function of *BRCA2* is in homologous recombination (HR) in which the BRCA2 protein binds RAD51 protein at its binding domain, which is a region of eight BRC repeats. RAD51 protein is an enzyme that instigates the exchange between homologous DNA molecules via the formation of a nucleoprotein filament that covers

ssDNA molecules. Essentially, *BRCA2* regulates *RAD51* recombinase (Venkitaraman 2002).

Figure 2.2 The *BRCA1* network in response to DNA damage

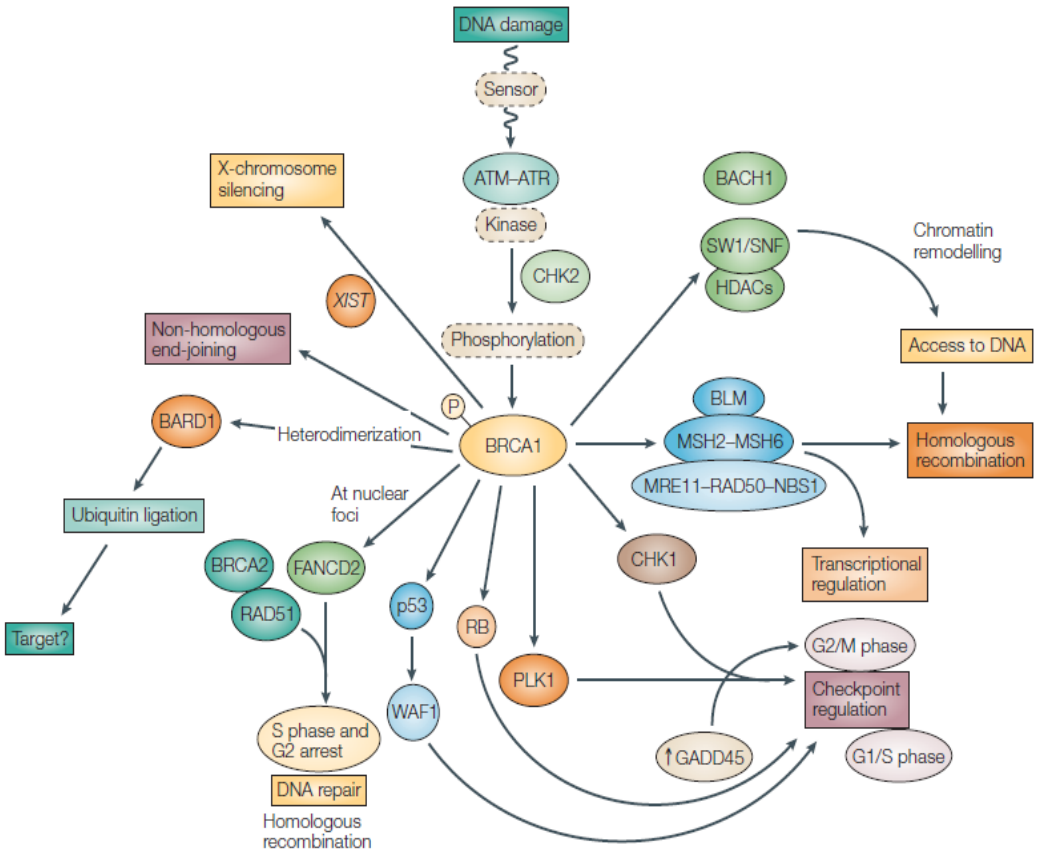


***Figure 2.2. The BRCA1 Network in Response to DNA Damage**. Adapted from: Narod, S.A. & Foulkes, W.D. (2004) Nat.Rev.Cancer, (4):665-676.*

2.4 Frequency of *BRCA1* and *BRCA2* mutations in specific populations

There are large variations in the frequency of *BRCA1* and *BRCA2* mutations detected within specific populations. A founder mutation can be defined as one that persistently arises within one haplotype in a specific population (Fackenthal & Olopade 2007). A haplotype is a set of genetic markers (usually Single Nucleotide Polymorphisms or microsatellite markers) that are inherited together as a block. Haplotypes arise through meiotic recombination during which fragments of chromosomes are mixed. Interbreeding leads to the same fragments appearing in a number of different individuals. When recombination is random, these fragments would be dispersed over time (International HapMap Project 2011). Founder mutations may continue to be evident within the specific population or may become incorporated within a wider

community such as a particular geographic region due to historical relocation of the community in which the founder mutation occurred.

Ramus & Gayther (2009) report vast differences in the frequency of particular mutations detected in diverse populations. Founder mutations are evident in specific ethnic groups or in localised global areas. The presence of founder mutations is relevant in the genetic screening laboratory in that, in those populations containing a large percentage of founder mutations, screening could be specifically targeted for those identified mutations. For example, in the Ashkenazi Jewish population there are three identified mutations that explain most of the specific founder mutations recognised in this ethnic group. In this population the prevalence rate of one of three mutations in *BRCA1/2* is 2.6% compared to the general population rate of 0.1-0.2%. These three mutations are 185delAG (*BRCA1*), which has a prevalence rate of 1%; 5382insC (*BRCA1*) with a prevalence rate of 0.13 % and 6174delT (*BRCA2*), which has a prevalence rate of 1.52 % (Ferla et al 2007). Together these three mutations constitute 98-99% of all mutations within the Ashkenazi Jewish population (Phelan et al 2002). Thus, mutation specific screening for the Ashkenazi Jewish population is widely accepted in routine clinical use (Ramus & Gayther 2009). In specific regions or isolated countries around the world similar clinical screening is employed for other distinct founder mutations. For example, the Icelandic *BRCA2* mutation 999del5 causes most of the HBOC or familial ovarian cancer only cases attributable to *BRCA1* or *BRCA2* genes.

In Iceland the most prevalent founder mutation is in *BRCA2* gene (995del5), which occurs at a rate of 0.4% of the population. However, this mutation occurs at a rate of 7.9% in breast cancer patients and at a rate of 8.5% in ovarian cancer patients. Also in Iceland is the more rare, but still significant founder mutation G1593A in *BRCA1*, which is identified in 1% of breast and/or ovarian cancer patients. The Table 2.1 below defines the key founder mutations within Europe in *BRCA1* and *BRCA2* genes.

2.4.1 The key founder mutations

Table 2.1 Key founder mutations in European populations

Name of Population	*BRCA1*/*BRCA2* mutation	Prevalence in named population (%)	Contribution to *BRCA1*/*2* mutations
Ashkenazi Jewish	*BRCA1* 185delAG	1.00	16-20% breast cancer diagnosed under 50 years
	BRCA1 5382insC	0.13	
	BRCA2 617delT	1.52	8% women diagnosed with breast cancer under 42 years and 7% women diagnosed with breast cancer 42-50 years that have a strong family history
Icelandic	*BRCA1* G5193A	rare	1% breast/ovarian cancer diagnoses
	BRCA2 995delT	0.4%	8.5% breast cancer patients and 7.9% ovarian cancer patients
Norwegian	*BRCA1* 1675delA *BRCA1* 816delGT *BRCA1* 3347delAG		These 3 account for 68% of all the BRCA1/2 mutations
	BRCA1 1135insA		Found in 3% of ovarian cancer
Finnish	*BRCA1* IVS11+3A>G *BRCA2* 9345+1G>A *BRCA2* C7708T *BRCA2* T8555G		These 4 mutations constitute 84% of all BRCA1/2 mutations
Swedish	*BRCA1* 317ins5		70% of BRCA1/2 mutations in one region (west of Sweden)
French	*BRCA1* 3600del11 *BRCA1* G1570X		52% of BRCA1/2 mutations
Dutch	*BRCA1* 2804delAA *BRCA1* IVS12-1643del 3835 *BRCA2* 5579insA *BRCA2* delTT		All 4 mutations constitute 24% of BRCA1/2. Both BRCA2 mutations constitute 62% of those with strong family history
Italian (Calabria)	*BRCA1* 5083del19		
Italian (Sardinia)	*BRCA2* 8765delAG		1.7% in breast cancer patients
Polish	*BRCA1 5382insC* *BRCA1 C61G* *BRCA1 4153delA*		*Sourced from Gorski et al (2004)*

Table 2.1 Key founder mutations in named European populations*. This table describes the key founder mutations in European populations, including prevalence within breast/ovarian cancer patients, if known. Table adapted and re-drawn from Ferla et al (2007). Annals of Oncology 18 (Supplement 6) vi93-vi98.*

Table 2.2 Key founder mutations identified in non-European populations

Name of population	*BRCA1*/*BRCA2* mutation
French-Canadian	*BRCA1* C4446T *BRCA1* R1443X *BRCA2* 8756delG *BRCA2* 3398delAAAAG
Hispanic (South Carolina)	*BRCA1* S995X *BRCA1* 2552delC
Hispanic (Columbia)	*BRCA1* 3450delCAAG *BRCA1* A1708E *BRCA2* 3034delACAA
African-American	*BRCA1* 943ins10 *BRCA1* 1832del5 *BRCA1* 5296del4 *BRCA2* IVS13 + 1G>A
South African	*BRCA1* E881X
Jewish (Iraq/Iran)	*BRCA1* Tyr978X
Chinese	*BRCA1* 1081delG
Japanese	*BRCA1* Q934X *BRCA1* L64X *BRCA2* 5802delAATT
Malaysian	*BRCA1* 2846insA
Filipino	*BRCA1* 5454delC *BRCA2* 4265delCT *BRCA2* 4859delA
Pakistani	*BRCA1* S1503X *BRCA1* R1835X
Afrikaner	*BRCA2 8162delG (sourced from Schoeman et al 2013)*

Table 2.2 Key founder mutations identified in named non-European populations*. This table describes the key founder mutations in non-European populations. Adapted and re-drawn from Ferla et al (2007). Annals of Oncology 18 (Supplement 6) vi93-vi98.*

Screening for founder mutations is straightforward and raises the possibility of rapid population level targeted genetic testing. This would also be cost effective in that only the specific variants in certain populations are screened and allows for the possibility of extending genetic testing to specific populations rather than restricting testing to cases with a strong family history of the disease, which is currently the case. Using this approach more accurate prevalence and penetrance estimates can be made within these populations (Ferla et al 2007). Widening research to new populations, screening additional genes and examining for large rearrangements may uncover further founder mutations.

2.5 Mutations in *BRCA1* and *BRCA2* and ovarian cancer susceptibility

The predominant ovarian tumours are epithelial carcinomas. Of these the most common histological sub-type is high-grade serous carcinoma (Gilks & Prat 2009), which notoriously presents at a late stage when metastatic lesions are already evident.

The majority of high-grade serous carcinomas reveal aberrations in *BRCA1* or *BRCA2*, these alterations include, epigenetic changes as well as germline and somatic alterations. In addition, often p53 is mutated or lost. Genomic instability results from the non-repair of DNA damage (Gilks & Prat 2009).

2.5.1 Mutation detection in *BRCA1* and *BRCA2*

Germline mutations are distributed throughout *BRCA1* and *BRCA2* coding regions and are of various types, including small mutations, insertions and deletions that result in a frameshift (i.e. alteration of the gene's reading frame and a non-functioning protein) and nonsense mutations (i.e. that result in the addition of a stop codon and again a truncated non-functioning protein product). Thousands of pathogenic mutations have been found in *BRCA1* and *BRCA2* genes (Breast Cancer Information Core BIC). These mutations include protein-truncating mutations that result in a loss of protein function; for example, nonsense and frameshift mutations. A proportion of these variants are missense mutations, of which a small subset are deleterious, however, many are variants of uncertain significance (VUS). Interestingly, certain pathogenic mutations have been found in higher density in exon 11 of *BRCA2* and these reveal an elevated ovarian cancer risk compared to breast cancer. This region is now referred to as the ovarian cancer cluster region (OCCR) of *BRCA2* (Gayther et al 1997).

2.5.2 Mutation types

Genetic variation or changes in the DNA sequence can take many forms and may or may not be pathogenic. These changes may or may not result in an alteration of the amino acid. Pathogenic mutations can range in size and type, for example, point mutations (deletions or insertions) or larger genomic rearrangements including deletions, insertions and duplications. Different types of variants give rise to different consequences and may or may not affect gene expression. For example, the inactivation of tumour suppressor genes can result from missense, frameshift or nonsense mutations since these can lead to a truncated and non-functioning protein product. Frameshift mutations involve the addition or loss of base pairs that lead to an alteration in the gene's reading frame. As a result the grouping of bases is shifted so that the amino acid code is altered and ultimately means a non-functioning protein. Missense mutations are an alteration in one base pair that leads to the substitution of one amino acid for another. Nonsense mutations are an alteration of one base pair that lead to the insertion of a stop codon and therefore, a truncated non-functioning protein. Splice-site mutations occurring at splice sites have various consequences; for example, these can result in exon skipping; or the exploitation of cryptic spice sites

resulting in aberrant mRNA that results in the insertion of a stop codon and a truncated non-functioning protein product (Speicher et al 2010)

The first mutations to be identified in these two genes were protein-truncating mutations, mostly nonsense, small insertions or deletions. The Breast Cancer Information core (BIC) is a database, for *BRCA1* and *BRCA2,* of both type of mutation and their occurrence. Initially, the majority of mutations identified were protein truncating therefore, most research centred on the use of the protein-truncation test (PTT); however, this test does not identify small mutations, particularly those in small exons. The PTT will not identify mutations within the regulatory regions or intronic regions that may affect RNA stability (Narod & Foulkes 2004) or mutations that lead to protein product that does function albeit in an altered fashion.

Early attempts to detect mutations in *BRCA1* and *BRCA2* focus on the coding regions and splice site of both genes. This is undoubtedly due to the sequencing methods available at the time. These experiments identified many disease-causing mutations in both genes. However, these leave some gaps in the sequencing of the entire genomic regions of these genes in that the intronic and regulatory regions are yet to be sequenced and mutational analyses conducted.

Ramus et al (2007) investigate the contribution of *BRCA1* and *BRCA2* mutations to inherited ovarian cancer. This is achieved by analysing 283 epithelial ovarian cancer families for mutations in *BRCA1* and *BRCA2.* Various mutation detection methods are used, including Multiplex Ligation Probe Amplification (MLPA), Heteroduplex Analysis (HA) and capillary electrophoresis, to screen the coding regions and splice sites of both genes. They report the prevalence of mutations in *BRCA1* as 37% and in *BRCA2* as 9%. The predominant mutations they detect are frameshift and nonsense mutations with the majority of these residing in a central portion of each gene. This central portion of *BRCA2* has been known previously as the ovarian cancer cluster region (OCCR) (Gayther et al 1997). In the Ramus et al (2007) study they report that 85% of the detected *BRCA2* mutations are located in this 3.5 Kb central section of the gene. They conclude that overall the stronger the family history of breast/ovarian cancer, the higher the likelihood of *BRCA1* or *BRCA2* mutations.

Genetic linkage analysis is the examination of genetic markers that have been inherited together. Ramus et al (2007) examine genetic linkage to *BRCA1* or *BRCA2* loci, using microsatellite markers, in those epithelial ovarian cancer families that do not have mutations in either *BRCA1* or *BRCA2* genes. This is achieved by combining linkage analysis in the family and loss of heterozygosity (LOH) analysis in tumours.

Interestingly, 5/9 non-*BRCA1* or *BRCA2* families are shown to be linked to *BRCA1* or *BRCA2* loci, suggesting the possibility that mutations may be present that are missed by screening. However, as 4/9 families reveal no linkage to *BRCA1* or *BRCA2* mutations, this proposes the likelihood that other susceptibility genes exist.

In order to build on the work by Ramus et al (2007), using the latest next generation sequencing technology (NGS), these two genes can now be sequenced in their entirety, including all intronic and regulatory regions. NGS can also be used to sequence other candidate susceptibility genes identified from genome wide association studies or those involved in the DNA repair pathway. This may involve sequencing large numbers of cases and controls in ovarian cancer and/or breast cancer.

Morgan et al (2010) use NGS to sequence the coding regions of *TP53*, *BRCA1* and *BRCA2*. Long Range PCR is used as a target enrichment method and Illumina GAII to sequence the coding regions plus an additional 20nt either side of each exon for both genes. This is achieved by amplifying *TP53* in 2 fragments of 3,289 bp (exons 2-9) and 12,346 bp (exons 10-11); *BRCA1* and *BRCA2* are amplified in 22 fragments ranging in size from 1,221 bp to 5834 bp. This paper is useful in that it demonstrates the viability to translate this technology into the diagnostic clinical setting; however, it does omit regulatory and intronic regions and there is no linkage analysis alongside it to demonstrate if there are missed mutations or if there are additional susceptibility genes. The research being undertaken here could fill these gaps and answer questions vital in the translation of research into the diagnostic clinic. The research uses robust and reproducible experimental design, indicating that Long Range PCR (LR-PCR) is a good target enrichment strategy for smaller genomic regions. However, its relatively small cohort size of 55 breast cancer patients does not allow for the accurate assessment of the contribution of *BRCA1* to familial breast cancer. One final point is that it may be possible to use genomic enrichment with NGS (rather than LR-PCR and MPS) to detect large genomic rearrangements, including larger deletions and insertions by analysing read depth of coverage (Yoon et al 2009). This will make the translation of this technology to the diagnostic setting more feasible as it will not be necessary to use additional techniques such as MLPA to detect larger genomic rearrangements.

2.6 Target enrichment strategies

2.6.1 Long Range PCR

Target enrichment essentially isolates and amplifies the genomic region of interest. There are a number of target enrichment methods available and each has advantages and disadvantages. These methods include Long Range PCR (LR-PCR), Molecular Inversion Probes (MIP) and Hybrid Capture.

Long Range Polymerase Chain Reaction (LR-PCR) is a modification of PCR and has been one of the most commonly used techniques as it does not require expensive equipment and is suited to all the various massively parallel sequencing platforms. Polymerase Chain Reaction (PCR) is a method for amplifying a target DNA template. The technique takes the form of three cycles of reactions. First, DNA denaturation, the separation of double stranded DNA (dsDNA). Secondly, primer annealing, the binding of each primer to each end of the template strand revealed in the denaturation step. Finally, the extension cycle, during which a complimentary strand of nucleotides are synthesised by a thermostable *Taq* polymerase using dexoxynucleotidetriphosphates (dNTPs). This results in amplification of the original template DNA sequence in an exponential fashion (Moody 2007). In LR-PCR large amplicons are amplified; achieving this requires the use of another enzyme, in addition to the *Taq* polymerase, which is a high-fidelity polymerase with proof-reading 3'-5' exonuclease ability in order to accurately amplify amplicons of ~10 Kb.

There are numerous advantages in using PCR, including the uniformity of coverage with the use of overlapping long PCR amplicons. PCRs of ~10kb in length are probably the maximum size to ensure a robust PCR product (Mamanova et al 2010). Another advantage is there is no need for high tech equipment, just a few primers is all that is required for each different PCR reaction. Primers are inexpensive, easy to acquire and PCR is well established in the laboratory.

LR-PCR is not without its caveats, and these need to be addressed during experimental design. LR-PCR is likely to be most efficient in amplifying amplicons up to 10 Kb in size. Therefore, it is probably only realistic to amplify targets up to ~200 Kb; much more than this would require a large number of PCR reactions and would result in inflated costs due to the amount of primers and other reagents required as well as the DNA input requirements for larger numbers of reactions. To guarantee full coverage of the desired region and to ensure that primer annealing is not in areas of SNPs, primers are designed so that they amplify overlapping PCR fragments of several

hundred base pairs. Normalisation of amplified PCR products is necessary because, all reactions will not result in the same yield, even under the same conditions with identical concentrations of starting material (Mamanova et al 2010).

LR-PCR and the alternative methods for target enrichment can be assessed on several criteria, including specificity, or a measure of the amount of sequenced reads accurately mapping to the target region; sensitivity, illustrates the amount of bases covered by sequence reads; coverage uniformity, describes how consistent read depth is across the target region; reproducibility, assesses how well similar results can be achieved from repeat experiments; cost effectiveness; and input DNA requirement. Some of these parameters are interrelated in that a target enrichment method that shows both good coverage uniformity and specificity will be cheaper, since it would require reduced sequencing capacity. Issues that need to be considered when identifying the most appropriate approach to target enrichment include the size of the target region and the quantity of samples to sequence (Mamanova et al 2010).

2.6.2 Molecular Inversion Probes (MIP)

Molecular inversion probes are an enzymatic solution to target enrichment. Oligonucleotides are constructed that are made up of a linker sequence and a target specific sequence on either side that then hybridise to the target region after which an extension step fills the gap between the two targets. Finally, a ligation step circularises the oligonucleotides (Figure 2.3). This method is useful for capture of exons or few regions in which a large number of samples are to be sequenced, since scaling up is relatively easy.

Figure 2.3 Molecular inversion probes

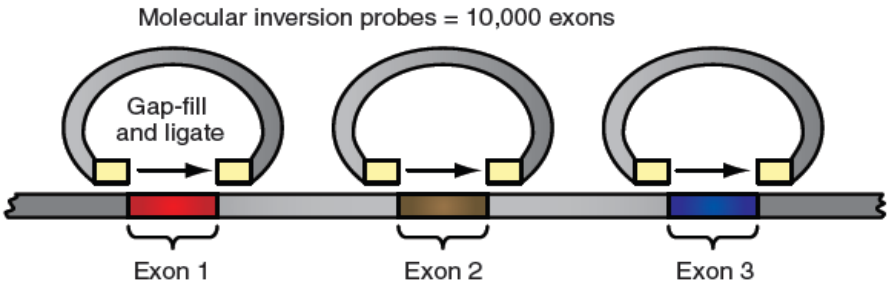


Figure 2.3 Molecular inversion probes. *Adapted from Mamanova et al (2010) Target enrichment strategies* for next generation sequencing. Nat. Methods. Vol 7, No. 2.**111-118.**

MIP demonstrate relatively poor coverage uniformity, but perform better on other parameters such as an input DNA requirement of as low as 200ng and sensitivity of more than 98%.

2.6.3 Hybrid capture

Hybrid capture is a genomic sequence capture method developed by Roche Nimblegen and Agilent (Figure 2.4). It can be performed on array or in solution. In this method the DNA libraries are prepared first from genomic DNA, which is then annealed to probes specific to the target region. The DNA that is not specific to the target sequence is simply washed away and the captured target DNA is eluted for sequencing.

Figure 2.4 Hybrid capture on array or in solution

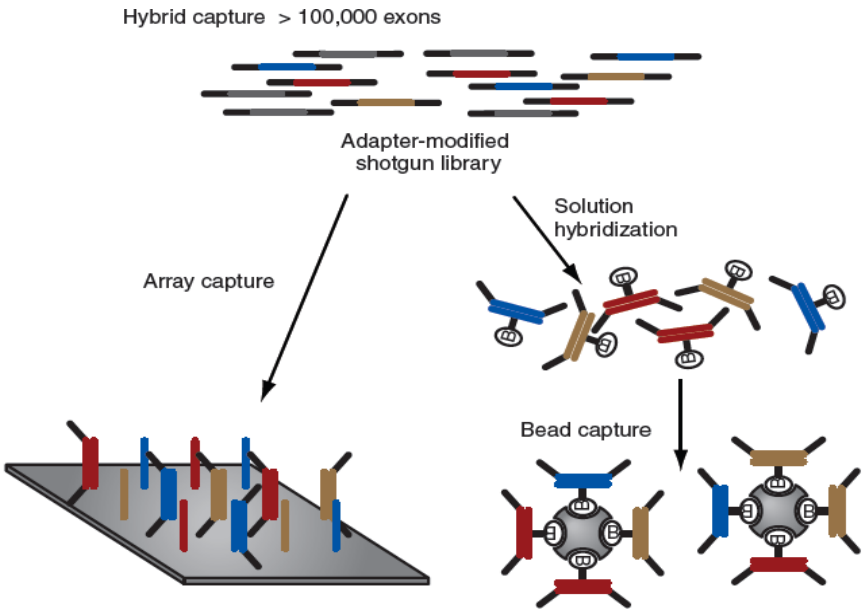


Figure 2.4 Hybrid Capture on array or in solution**. Adapted from Mamanova et al (2010) Target enrichment strategies for next generation sequencing. Nat. Methods. Vol 7, No. 2.**111-118.

The obvious advantage with this method of target enrichment is in its speed and ease of use, however coverage uniformity does not compare well with LR-PCR, at ~60%. Other parameters are good with small quantities of input DNA requirements and high sensitivity. Specificity is below that achieved by the other methods at ~70-80% (Mamanova et al 2010).

Whilst PCR is probably not suitable for larger regions, it is a highly specific method that can offer good coverage uniformity at a relatively low cost for smaller target regions. If, however, the whole exome is required alternative methods of target enrichment, such as hybrid capture are more appropriate.

2.7 The Illumina Genome Analyser II (GAII) Platform

Figure 2.5 Illumina GAII flow cell

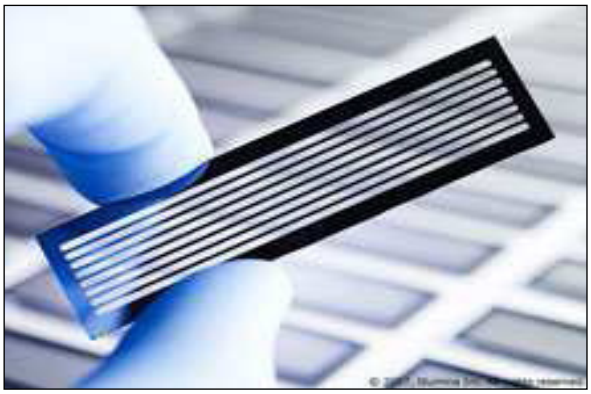


Figure 2.5 Illumina GAII Flow Cell. *This picture shows the Illumina GAII flow cell, which consists of a solid substrate, made of glass and silicon, with 8 discrete channels on to which millions of oligonucleotides are hybridised. These oligonucleotides create a lawn across the surface of the flow cell and act as probes to which the prepared DNA fragments hybridise. (Image from www.illumina.com).*

2.7.1 Library Preparation

Figure 2.6 Library preparation

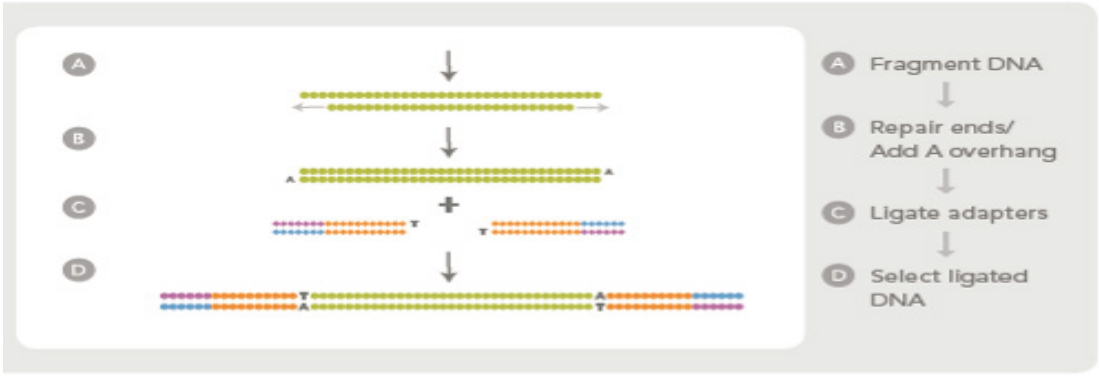


Figure 2.6 Library preparation*. This flow diagram shows the steps involved in library preparation from genomic DNA to prepared libraries ready for sequencing on the GAII. (Image from www.illumina.com).*

Sequencing libraries are prepared; one library for each sample analysed. This is done by first, fragmenting the template DNA into ~200-600bp sized fragments (A. Figure 2.6) via one of 2 possible methods (nebulisation or sonication). The template DNA used is the pool of LR-PCR fragments for each sample generated during target enrichment. Next the fragments are denatured and the ends of fragments repaired to form blunt ended fragments (B. Figure 14). Then an 'A' base overhang is added (B. Figure 2.6), which allows for adapter sequences (complimentary to the oligonucleotides on the flow cell surface) to be ligated (C. Figure 2.6).

If multiplexing, the 6 base index sequences are introduced at the PCR enrichment stage, during which the selected DNA fragments are enriched (D. Figure 2.6). For a paired end (PE) read there are two primers for the PE read (one forward and one

reverse), plus one primer for the index. These can be added to both samples and to genes, so that all of *BRCA1* can be sequenced for multiple patients in one lane of one flow cell. During this pilot study Illumina supplied 12 index primers that contain no more than 3 positions the same between each index so that if there is an error in one the correct index can be identified. These 12 indexes can be used for each lane of the flow cell. By the end of this pilot study Illumina increased multiplexing levels to 96 index sequences.

2.7.2 Cluster Generation

Figure 2.7 Cluster generation

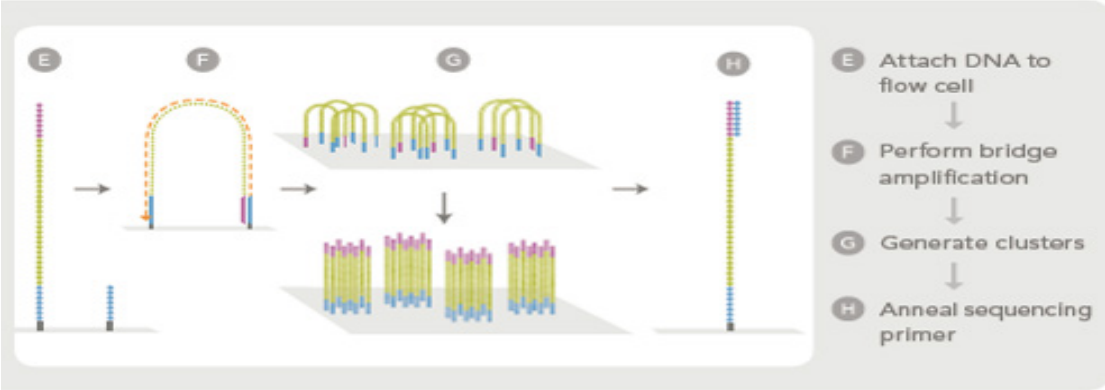


Figure 2.7 Cluster generation*. This flow diagram describes the process of cluster generation of prepared libraries prior to sequencing on the GAII. (Image from www.illumina.com).*

Prior to sequencing, the prepared templates must be copied. The cluster station (recently upgraded to the 'CBot') is the separate fluidics apparatus that conducts this step. The fragmented DNA is first denatured into single stranded fragments and attached to the surface of the flow cell (E. Figure 2.7). A new strand is created with polymerase by extending the template strand. The adapter sequences of the newly synthesised strand, of which there are two different ones (forward and reverse) arch over and anneal to a free oligonucleotide on the flow cell (F. Figure 2.7) This produces a bridge and a new position for the synthesis of another new strand. Forward and reverse fragments are generated by repeating the process and produce 100s of millions of clusters each with ~1,000 copies of the original template (G. Figure 2.7). These clusters are denatured and cleaved to leave just the forward strand (for single read sequencing) allowing for many simultaneous sequencing reactions, hence the term 'massively parallel'. Primers for sequencing are hybridised to templates and the flow cell is moved to the GAII (H. Figure 2.7)

2.7.3 Sequencing-by-synthesis (single read sequencing)

Figure 2.8 Sequencing-by-synthesis

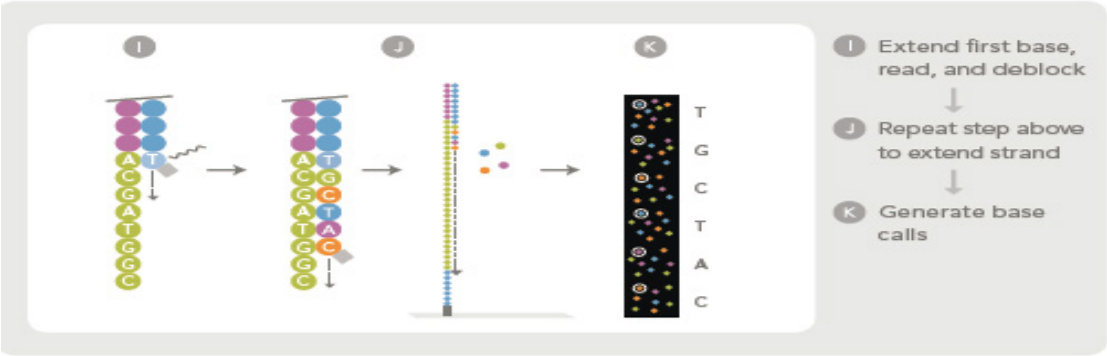


Figure 2.8 Sequencing by synthesis. *This image describes sequencing by synthesis using reversible chain termination chemistry. (Image from www.illumina.com).*

Once the clusters are generated, the flow cell is moved to the GAII where the clusters are first denatured. Then the polymerase, primers for sequencing and modified fluorescently labelled chemically modified nucleotides (reversible dye terminators) are included. It works as follows: the 4 fluorescently labelled nucleotides are modified so that the 3'OH can be chemically inactivated to permit reversible chain termination. As each base is synthesised and the chain terminated due to the inactivation of the 3'OH in the nucleotide, this base is read and the blocking of the 3'OH is removed to allow for the next base to be synthesised. The process is repeated and the flow cell surface is imaged following the addition of each base (Tucker et al 2009). Each of the fluorescent bases is detected via laser excitation.

2.7.4 Paired-end sequencing

Paired-end sequencing requires specialised additional equipment from Illumina, known as the paired-end module that is connected to the GAII. The Paired-End Module is an additional piece of specialised equipment that is attached to the Genome Analyser.

Figure 2.9 demonstrates how paired end sequencing is achieved; essentially both ends of the DNA template are sequenced and paired with genomic inserts of a known size. In the first read the reverse strand is re-synthesised from the forward strand, the forward strand is subsequently cleaved to allow for complimentary strands to be bridge amplified creating new clusters for the second read. Sequencing primers SP1 and SP2 are paired and sequencing-by-synthesis is performed in order.

Figure 2.9 Paired end sequencing flow diagram

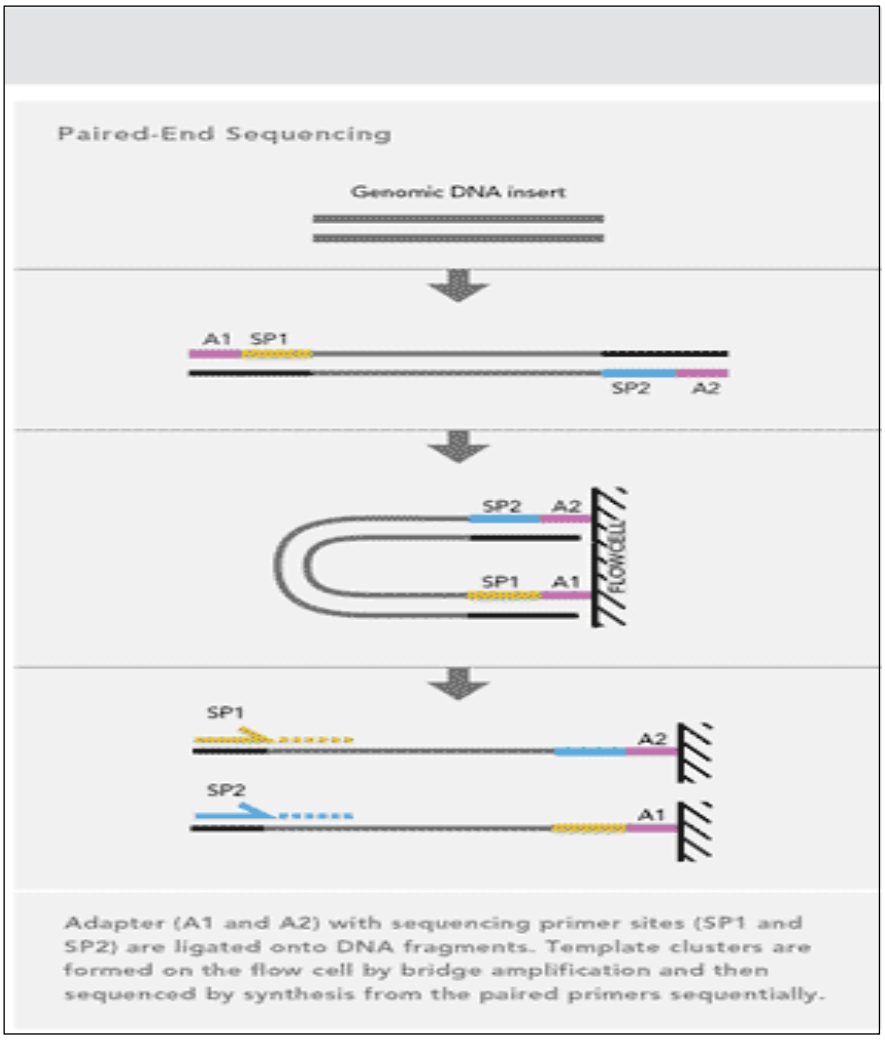


Figure 2.9 Paired end sequencing flow diagram*. This diagram demonstrates how paired end sequencing is achieved, so that each DNA template is sequenced from each end in the forward and reverse directions. (Image from www.illumina.com).*

2.7.5 Multiplexed Sequencing

The expected coverage on a 76bp single-end read for one channel of the flow cell is ~30,000X. This calculation is made using the Lander & Waternan (1988) formula: C = LN/G where C = coverage, L = read length (bp), N = number of reads and G = haploid length of the genome (bp). So, C = 76 x $38x10^6$ / 91,500 = 31,562X. Thus, it is logical to put this large sequencing capacity to full use. Multiplexing using bar-coded index sequences enables several samples and several genes to be sequenced in one lane, improving efficiency by lowering the cost per sample as well as reducing the time taken to sequence several genes. To sequence 12 patient samples for two genes would give coverage of ~1,250X. One issue that arises when pooling samples and LR-PCR amplicons is the possible loss of sequencing uniformity. Multiplexed sequencing is possible via the addition of tags (index sequences) enabling 12 samples to be

sequenced in each lane or 96 samples per flow cell. The next diagram shows how this is achieved (Figure 2.10).

Figure 2.10 Multiplexed sequencing

Figure 2.10 Multiplexed sequencing*. Multiplexing is achieved via the addition of a 6 base index sequence at the PCR enrichment stage following ligation of adapter sequences as the final step in library preparation. (Image from www.illumina.com).*

2.8 Bioinformatics for data analysis and mutation detection

2.8.1 Data analysis of DNA sequencing using Capillary Electrophoresis

Sequence analysis for capillary electrophoresis is performed using either Applied Biosystems own program, SeqScape® or alternative programs that are available, for example, GeneScreen that has been developed by Ian Carr at the Leeds Institute of Molecular Medicine, University of Leeds (Carr et al 2011). Both software programs have been designed to rapidly and accurately analyse capillary sequencing traces for variant detection or SNP analysis. These methods assist in improving throughput for data analysis.

To enable the identification of sequences produced by capillary sequencing freely available on line software programs are available to align sequences. The NCBI tool

BLAST (Basic Local Alignment Search Tool) is one such tool that searches for sequences with homology to the input template sequence.

2.8.2 Data analysis for DNA sequencing using Illumina GAII and HiSeq2000

The data analysis pipeline for Illumina GAII can be divided into a number of separate steps. These include: demultiplexing of indexed samples (for multiplex sequencing), read mapping (alignment of reads to the reference sequence), base calling and variant detection. Bioinformatics analyses also include additional statistical information such as quality scores and depth of coverage.

Data analysis of NGS data is a rapidly developing field with many modifications of existing approaches as well as new approaches emerging at an alarming rate. Most of these modifications centre on the algorithms in use to align the multitude of short read sequences produced by MPS technology. Differences in these can result in differences in data output.

2.8.3 Sample Demultiplexing

If multiplexed sequencing is conducted, in which several samples are sequenced in each lane of the flow cell, the first step in data analysis is the demultiplexing of indexed reads to accurately assign reads to each sample. The index sequences have no more than 3 bases that are the same between each of them; this ensures that if there is an error in one the correct index can still be identified and good quality reads do not get filtered out due to a mismatch in the bar code sequence.

The diagram, Figure 2.11, shows the workflow of the analysis of NGS sequencing data.

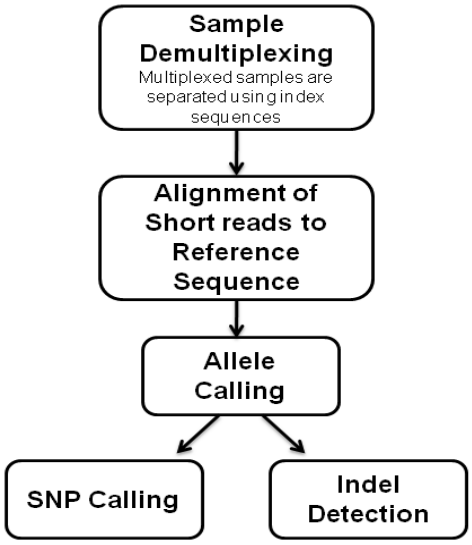


Figure 2.11 NGS sequencing data analysis workflow

2.8.4 Alignment of reads to the reference sequence (read mapping)

The next step in analysis is the alignment of millions of short sequencing reads to the reference sequence. This is possibly the most crucial step in NGS data analysis and the subsequent accurate detection of variants relies upon precise read alignment. The previous methods to align sequences to the reference sequence, including BLAST, are not appropriate for mapping the abundant 76bp reads produced by Illumina GAII. The GAII data will undoubtedly include gaps and these are dealt with by the clever use of algorithms. In addition, algorithms need to be robust enough to deal with the abundant repetitive sequences in the genome.

The Consensus Assessment of Sequence and Variation (CASAVA) is a package that includes read mapping using the mapping algorithm known as Efficient Large-Scale Alignment of Nucleotide Databases (ELAND). ELAND identifies the location on the reference sequence of sequencing reads. To do this ELAND divides a 32 bp read section (a 'seed') into 4 portions of 8 bp. These 8 bp sections are called 'substrings'. Substrings are then concatenated to form 32 bp seeds; with two mismatches permitted so that two (or more) of the 4 substrings will map to the reference sequence (Horner et al 2009). This should allow for the accurate alignment of 'gapped reads'.

Alternative programs that use modified algorithms have been developed for read mapping. The differences in these include, speed, amount of memory space required and the accuracy by which indels can be detected. Size of allowable indels varies from program to program (Horner et al 2009).

The Table 2.3 below outlines some of the available software programs for read alignment and variant detection. Selecting which program to use will often depend on the available computer systems in terms of the memory capacity and whether gapped or ungapped read alignment is required. In the case of variant detection, as opposed to *de novo* assembly, gapped read alignment is vital.

Table 2.3 A selection of the available software for NGS data analysis

Program Name	Description	Reference/Website
Bowtie	Aligns reads to reference sequence – low memory requirements	www.bowtie-bio.sourceforge.net
CLC Genomics Workbench	Complete integrated system including read alignment to reference sequence, SNP/indel detection, sequence viewer and statistical output	www.clcbio.com
ELAND	Gapped read alignment program. Developed for Illumina sequencing platform probably one of the fastest alignment programs with economical memory requirements	www.bioinfo.cgrb.oregonstate.edu/docs/solexa
Exonerate	Aligns reads to reference sequence	www.ebi.ac.uk/~guy/exonerate
MAQ	Read alignment is fast but lacks the accuracy of other programs	www.maq.sourceforge.net/index.shtml
MOSAIK	Gapped read alignment	
NextGENe	Gapped read alignment as well as SNP/indel detection and alignment viewer.	www.softgenetics.com
NOVOALIGN	Gapped read alignment	www.novocraft.com/products.html
RMAP	Read alignment	http://rulai.cshl.edu/rmap
SHRiMP	Read alignment	http://compbio.cs.toronto.edu/shrimp
SOAP	Alignment program – fast and accurate but requires a large amount of memory capacity	www.soap.genomics.org
ZOOM	Similar to ELAND, faster – but requiring more memory capacity	

***Table 2.3 A selection of the available software for NGS data analysis**. Information sourced from Horner et al (2009) Briefings in Bioinformatics vol 11. No 2 181-197.*

2.8.5 Base calling and variant detection

Once reads are mapped to the reference sequence those that match are stored as the 'Final Read Set' (CASAVA); one or two allele calls are made by the allele caller with a quality score (i.e. PHRED score divided by 10) assigned to each allele call. For a SNP to be called homozygous no reference allele must be seen; for a SNP to be heterozygous the second allele must score ≤ 6 (i.e. PHRED score 60) and the ratio between the two alleles must be ≤ 3 (i.e. PHRED score 30). Indels are detected by assembling reads that align; those that do not align fully (known as 'shadow reads') are assembled into contigs. Assembled contigs are aligned to the reference sequence again to produce candidate indels, which are compared to the reference to determine homozygosity, heterozygosity or no indel (CASAVA v 1.6 User Guide).

2.9 Clinical genetic screening for rare high-penetrance ovarian cancer and breast cancer susceptibility genes *BRCA1* and *BRCA2*

There are three main reasons for genetic testing for mutations in *BRCA1* and *BRCA2*. Firstly, for the assessment of an individual's cancer risk, secondly, to initiate a strategy for early prevention, including, prophylactic therapy (chemoprevention or risk reducing surgery) and early monitoring and thirdly, for the use of targeted treatments for patients with a cancer diagnosis. For example, cancer patients positive for *BRCA1* or *BRCA2* mutations have been found to be highly sensitive to PARP inhibitor drugs (Farmer 2005). These treatments are likely to be less toxic and more specific than the current chemotherapies.

The existing recommendations for genetic screening for *BRCA1* and *BRCA2* in the clinic are to test women for mutations if they have a 20% or more assessed risk of having a mutation (NICE 2006). Therefore, testing is currently only offered to women assessed as having a very strong family history. In addition, women with a strong family history of breast cancer or ovarian cancer must have either, a living affected relative that can be tested for mutation status; or be a patient with ovarian cancer or breast cancer to be offered genetic screening. Research has shown that a large proportion of women from low-risk families are positive for mutations (Walsh et al 2010); therefore, these patients would be missed in screening. Currently, one of the main reasons why a wider population of women are not tested is because the cost of screening makes it prohibitive.

2.9.1 Advantages and disadvantages of *BRCA* genetic testing

The main advantages of *BRCA* genetic testing include the identification of women at high risk of developing breast or ovarian cancer, which will allow for the introduction of targeted strategies for cancer prevention and early detection. It may also be advantageous to identify those that are not carriers of *BRCA* mutations; and this may relieve anxiety in those with a strong family history of disease. However, *BRCA* testing is not without drawbacks; it is likely that not all mutations will be detected, resulting in false negatives. The abundant variants of uncertain significance that are identified in these genes may in fact increase anxiety in individuals. Those patients diagnosed as negative for *BRCA* mutations would still be at risk from sporadic cancers. Some of the interventions for those diagnosed as *BRCA* mutation carriers, such as early screening, have not been adequately proven. Potentially, there could be a level of social or financial harm in having a test for a gene mutation as this may affect mortgage or life insurance applications.

When the diagnosis of a *BRCA* mutation is made the patient must make a decision on what measures to take; these often include screening and/or risk reducing surgery. In pre-menopausal women annual breast MRI scans are offered and possibly mastectomy. In *BRCA* mutation carriers there is also the option for pre-implantation genetic diagnosis (PGD) to test for *BRCA* mutations in embryos prior to in-vitro fertilisation (IVF).

The current method for genetic testing of *BRCA1* and *BRCA2* is PCR of specific exons and capillary electrophoresis to sequence for mutation detection. In addition, the use of multiplex ligation probe amplification (MLPA) for the detection of large genomic rearrangements. The regulatory regions and intronic non-coding regions are not currently included in routine testing. The cost for NHS clinical mutation testing in *BRCA1* and *BRCA2* during this study (conducted in 2010) is ~£700-£1,000, including all the coding regions of both genes and MLPA for large rearrangements. This cost has since reduced to ~£500 in 2014.

Genetic testing will be extended to those with an ovarian cancer or breast cancer diagnosis without family history and those with a strong family history without a cancer diagnosis, when the cost of mutation detection significantly reduces. A clinical genetics Department in Leeds suggest they can test the coding regions of *BRCA1* and *BRCA2* using LR-PCR and MPS for EU62.5 (~£50) (Morgan et al 2010). However, this cost does not include the intronic and regulatory regions. A group in US report that it cost US$1,500 (~£1000) to sequence multiple genes in their entirety using genomic capture and MPS and they suggest that this cost could be reduced to $500 if using multiplexed protocols (Walsh et al 2010). Using the protocols outlined in this report, Long PCR and MPS and multiplexing at 96 samples in one lane of a flow cell, it could possible re-sequence the whole of three genes for ~£320. The 'Discussion' section gives a full breakdown and comparison of costs for different sequencing methods in both research and clinical settings at the time of writing in 2010.

2.10 Identification of additional rare moderate-penetrance gene variants using NGS

Although, *BRCA1* and *BRCA2* are the two predominant high-penetrance genes in ovarian and breast cancer susceptibility these do not explain all inherited ovarian cancers and breast cancers. Five moderate-penetrance genes are identified as linked to a genetic predisposition to breast cancer, *ATM*, *BRIP1*, *CHK2*, *PALB2* and *NBS1*. These are detected by sequencing candidate genes for pathogenic mutations. It is highly plausible that these are also linked to epithelial ovarian cancer and that there are

other cancer susceptibility genes in this category, which could be successfully identified using NGS to sequence candidate genes in large cohorts of ovarian cancer cases and controls. These genes may also be identified with whole exome sequencing (i.e. sequencing the complete set of coding regions) in affected families.

2.11 Research aims

To conduct a pilot study screening for known mutations in 12 *BRCA1* mutation positive cancer patient DNA samples and demonstrate that Long Range PCR and NGS are as accurate as the current method (Sanger sequencing).

1. To conduct a pilot study that will include screening for mutations in regulatory and intronic regions of *BRCA1.*

2. To establish that, in principle, multiplexed NGS technology can be used to detect gene variants that predispose to an increased risk in ovarian cancer or breast cancer.

3. To conduct an assessment of the viability of Long Range PCR and NGS as a clinical screening method. This assessment will examine cost and time efficiency.

The pilot study aims to establish the protocols for multiplexed mutation detection, which will allow for the scaling up in follow on studies of larger cohorts of patients in which additional susceptibility genes will be analysed.

2.12 Hypotheses

1. NGS technology detects all known germline *BRCA1* variants in 12 mutation positive ovarian cancer patient control DNA samples.

2. It is hypothesised that NGS technology could be used as a clinical tool to detect gene variants that predispose to an increased risk in breast cancer or ovarian cancer.

3. This technology could be employed to identify rare variants in other genes that confer a moderate increase in cancer risk.

4. NGS is scalable for high-throughput research studies.

2.13 Results

(*Refer to Chapter 6 Materials and Methods page 250*)

Long Range PCR (LR-PCR) is used as a target enrichment method to amplify the whole genomic region of *BRCA1* gene. 11 overlapping primer pairs are designed using GenBank sequence L78833.1 Homo sapiens BRCA1 (*BRCA1*) gene, complete coding sequence. 12 DNA samples with known mutations in *BRCA1* are prepared for sequencing on the Illumina GAII. Several commercially available DNA polymerases are tested to find the best performing enzyme. One library is prepared for each sample and these samples are pooled in equimolar quantities and indexed with one of 12 unique barcode sequences. 76bp single read sequencing is performed on the GAII. Resulting data are analysed three times using different methods to identify all known mutations in the study DNA samples.

2.13.1 DNA samples

12 DNA samples are identified that have a variety of known mutations in *BRCA1* from cancer patients with breast, ovarian or prostate cancer. These mutations are not revealed to the author. This is done so that mutation detection can be conducted blinded to ensure good quality controls for the study.

2.13.2 The search for the best performing DNA polymerase for Long Range PCR

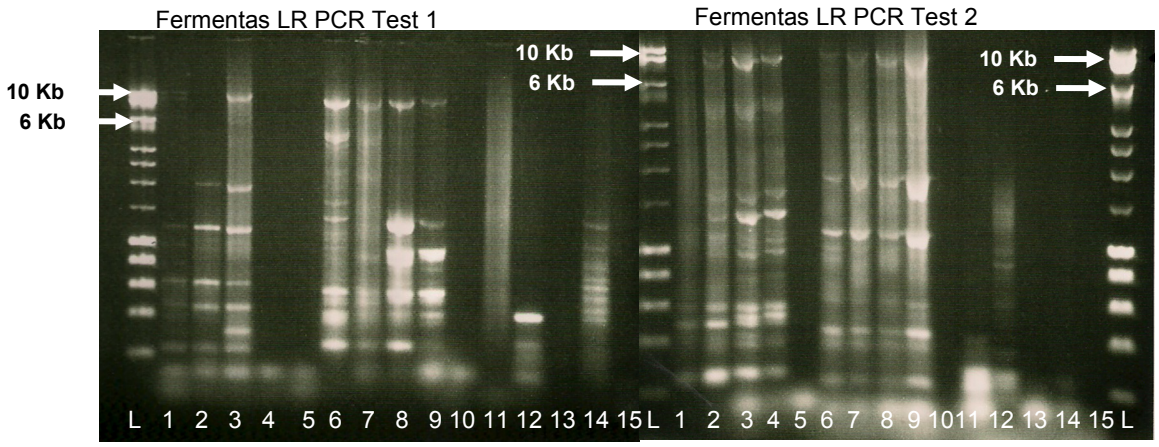


Figure 2.12 Fermentas Long Range PCR DNA polymerase. *Gel electrophoresis images of the same pooled control DNA sample using BRCA1 primers for one 6880 bp amplicon. Lanes 1-4, 6-9, 11-14 positive controls. Lanes 5, 10, and 15 negative controls.*

Figure 2.12 shows multiple bands and smearing for most lanes. Two tests are conducted on this enzyme using higher annealing temperatures on the second test.

Figure 2.13 Kapa Biosystems HiFi Hot Start Long Range

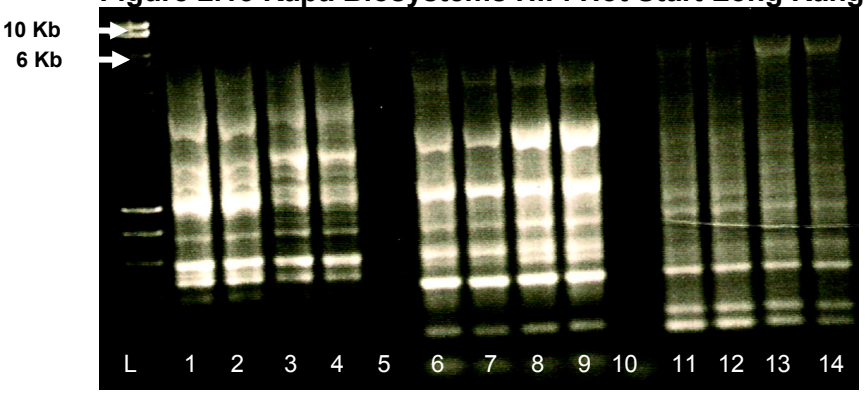


Figure 2.13 Kapa HiFi Hot Start Long Range. *Gel electrophoresis image of one example of 3 similar electrophoresis gels for three Kapa Biosystems enzymes tests using BRCA1 primers. Lanes 1-4, 6-9 and 11-14 positive controls; lanes 5 and 10 negative controls. This gel represents the PCR products of one 6880 bp amplicon in BRCA1*

Figure 2.13 shows multiple bands and smearing for all positive control lanes using the Kapa Biosystems range of DNA polymerases suitable for LR-PCR on a pooled control female DNA sample.

Figure 2.14 Phusion Hot Start Finnzymes

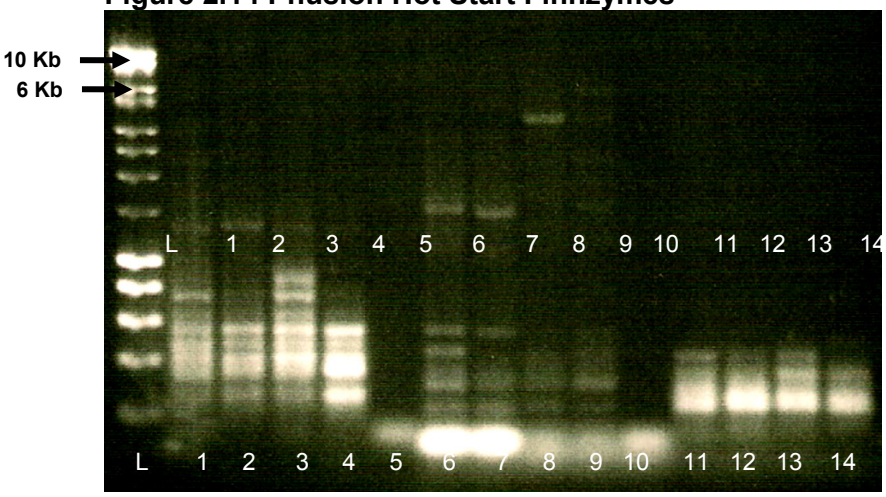


Figure 2.14 Phusion Hot Start from Finnzymes. *Gel electrophoresis image Phustion Hot Start Long Range PCR test using BRCA1 primers for one amplicon of 6880 bp. Lanes 1-4, 6-9 and 11-14 positive controls; lanes 5 and 10 negative controls.*

Figure 2.14 shows multiple bands and smearing for all positive control lanes using the Finnzymes enzymes Phusion Hot Start.

Figure 2.15 Invitrogen SequalPrep™ Long PCR Kit with dNTPs

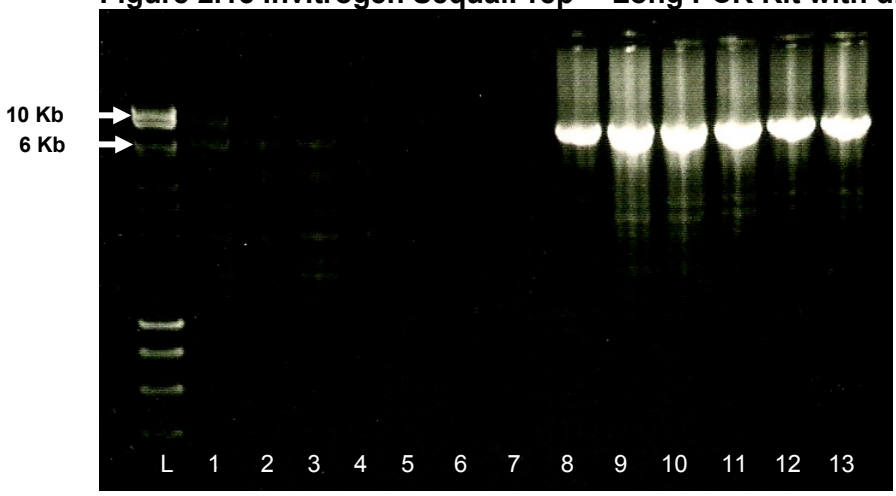


Figure 2.15 Sequal Prep Long PCR Kit with dNTPs. *Gel electrophoresis testing Invitrogen sequal prep Long PCR Kit. Lanes 1-6 positive controls using BRCA1 primers for 1 amplicon (intron 1 to intron 2); lane 7 negative control, lanes 8-13 positive controls using BRCA1 primers for 1 amplicon (intron 14 to 1)6. Lanes 8-13 show a PCR product of 6880bp.*

Figure 2.15 shows faint multiple bands for *BRCA1* primers intron 1 to intron 2 and clear bands for intron 14 to 16; lanes 11-13 show clearer and more specific bands using a higher annealing temperature. The SequalPrep Long PCR kit with dNTPs is chosen to amplify DNA samples from patients with known mutations in *BRCA1*, blinded to the author, to test the efficiency of this product on varying DNA samples. The Sequal Prep Kit is tested to see if DNA quality, assessed by age of the DNA sample, affects performance of the LR-PCR kit. 37 DNA samples, varying in quality, from both *BRCA1* mutation carrier samples and non-carrier samples are tested. Table 2.4 demonstrates these samples showing which reveal a good PCR result and which fail PCR.

Table 2.4 LR PCR enzyme efficiency and DNA quality

No	DNA Age in years	Sample ID	PCR Product Y/N	
1	23	OV001.306b	Y	12 carrier DNAs
2	19	OV002.304a	N	
3	19	OV002.304b	N	
4	19	OV025.515C	Y	
5	21	OV034.411a	Y	
6	16	OV034.411c	Y	
7	18	OV099.402a	N	
8	4	OV250.404	N	
9	11	OV371.304	N	
10	4	OV401.307	Y	
11	2	OV133.301b	Y	
12	2	OV069.301b	Y	
13	Degraded	PRY0861		Non- carrier DNAs
14	Degraded	PRY1145		
15	Degraded	PRY0777		
16	Degraded	PRS0625		
17	N/K	PRS0673	Y	
18	N/K	PRM4585	Y	
19	N/K	PRY1541	Y	
20	N/K	PRY0925	Y	
21	N/K	PRY3217	Y	
22	N/K	PRY3194	Y	
23	N/K	Pr_B1	Y	15 carrier DNAs
24	N/K	Pr_B2	Y	
25	N/K	Pr_B3	Y	
26	N/K	Pr_B4	Y	
27	N/K	Pr_B5	Y	
28	N/K	Pr_B6	Y	
29	N/K	Pr_B7	Y	
30	N/K	Pr_B8	Y	
31	N/K	Pr_B9	Y	
32	N/K	Pr_B10	Y	
33	N/K	OV089.305b	Y	
34	N/K	OV110.201c	Y	
35	N/K	OV110.303b	Y	
36	N/K	OV205.404	N	
37	N/K	OV362.403	N	

Table 2.4 LR-PCR enzyme efficiency and DNA quality. *This table shows that DNA quality appears to affect LR-PCR performance. NK= DNA age not known.*

The final samples chosen for sequencing are selected from those that produce a good LR-PCR product. It is important to note that DNA quality appears to affect LR PCR performance and that although the Invitrogen enzyme is a relatively robust product it performs best when using good quality DNA samples. A final set of 17 samples are chosen that work well for two fragments (6.436 Kb and 10 Kb), see Figure 2.17. Then these are amplified on all 11 fragments for *BRCA1*. The final 12 samples are chosen as those that amplified in all 11 fragments.

2.14 Target enrichment – Long Range PCR

The whole of each gene is amplified in 11 overlapping fragments. The diagrams below show where these fragments are located in each gene and where and by how many base pairs they overlap.

Figure 2.16 Long Range PCR amplification of *BRCA1*

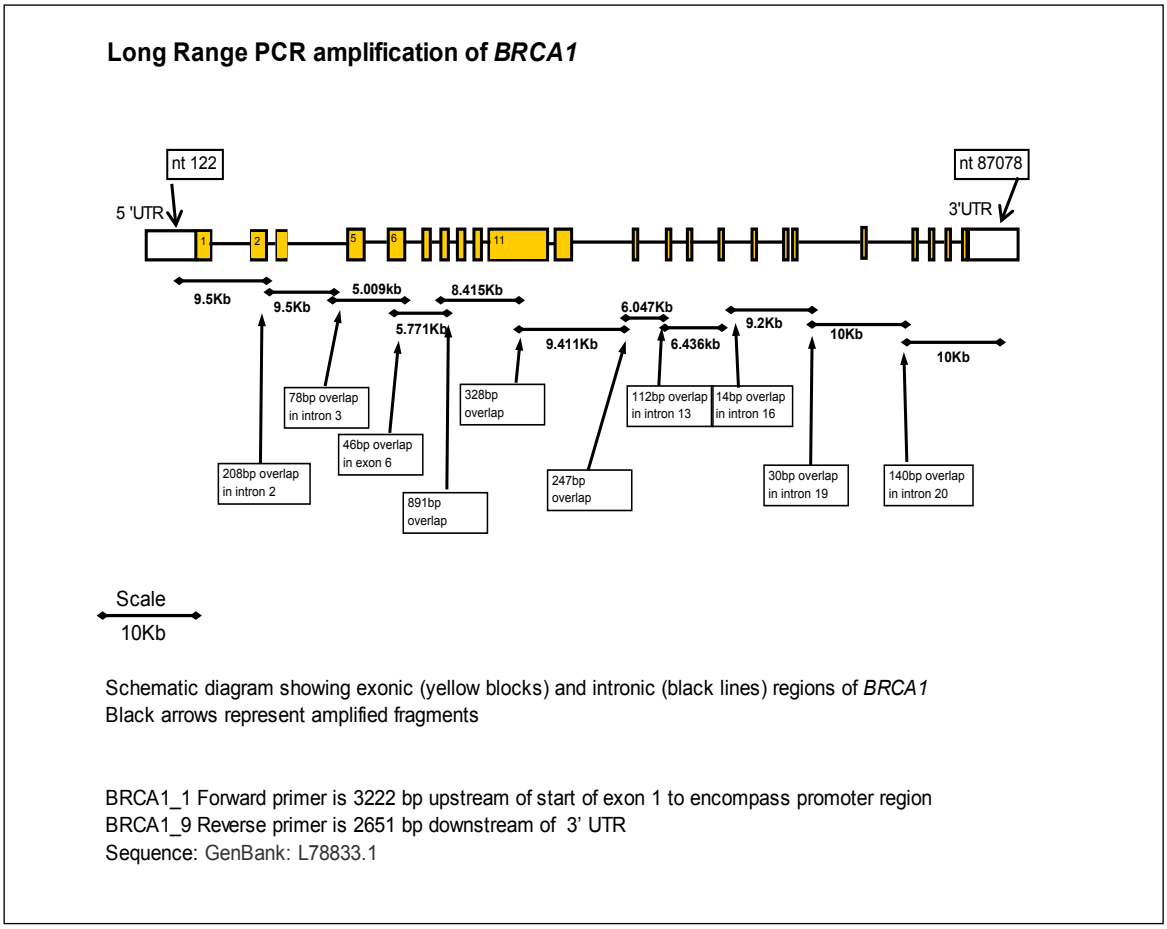


Figure 2.16 Diagram of Position and Size of Long Range PCR fragments that amplify the whole of BRCA1 gene.

Figure 2.17 Long Range PCR Gel Electrophoresis

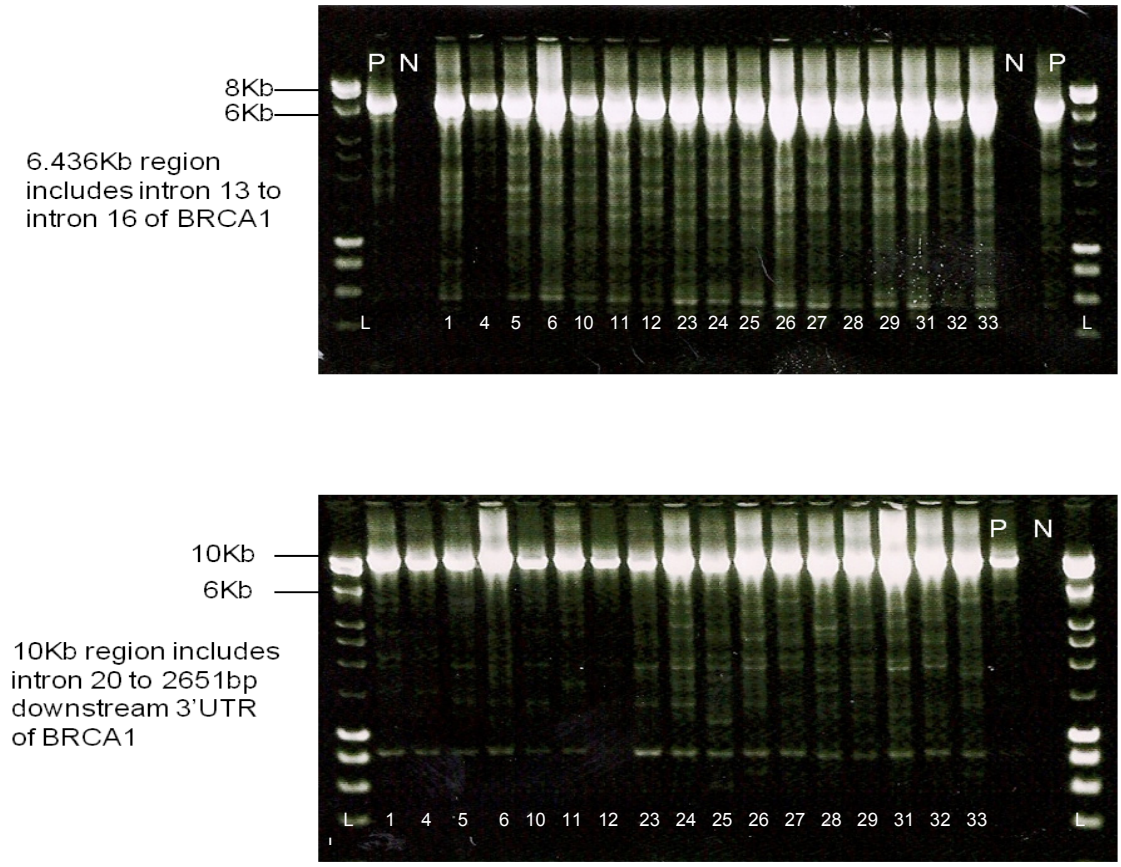


***Figure 2.17 Two representative examples of gel electrophoresis images for two fragments for 17 patient samples**. The final 12 samples are taken from these 17 samples. Lanes labelled P and N are Positive and Negative control lanes. All other lanes are patient DNA samples labelled with numbers corresponding to DNA sample numbers in Table 2.4*

Agarose gels are run for each fragment in 17 patient samples. The final 12 patient samples are selected from these. Figure 2.17 above are two representative gels.

2.14.1 PCR product normalisation

Table 2.5 Normalisation and pooling of PCR products

Pool 1				Pool 2				Pool 3			
Amplicon	ng/µl	DNA vol µl	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)
1	3.24	16.98	55	1	7.76	6.44	50	1	6.10	11.48	70
2	3.96	13.89	55	2	10.50	4.76	50	2	38.50	1.82	70
3	2.68	20.52	55	3	3.77	13.26	50	3	7.25	9.66	70
4	2.58	21.32	55	4	3.18	15.72	50	4	6.17	11.35	70
5	3.03	18.15	55	5	2.22	22.52	50	5	8.09	8.65	70
6	4.39	12.53	55	6	3.12	16.03	50	6	3.48	20.11	70
7	9.71	5.66	55	7	24.10	2.07	50	7	18.20	3.85	70
8	21.4	2.57	55	8	19.40	2.58	50	8	15.50	4.52	70
9	2.87	19.16	55	9	2.88	17.36	50	9	13.10	5.34	70
10	2.77	19.86	55	10	8.94	5.59	50	10	16.00	4.38	70
11	3.12	17.63	55	11	2.85	17.54	50	11	13.10	5.34	70
		168.3	605			123.87	550			86.50	770
Pool 4				Pool 5				Pool 6			
Amplicon	ng/µl	DNA vol	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)
1	3.85	14.29	55	1	6.39	10.95	70	1	5.59	8.94	50
2	19.30	2.85	55	2	21.40	3.27	70	2	11.80	4.24	50
3	8.96	6.14	55	3	3.35	20.90	70	3	2.45	20.41	50
4	5.29	10.40	55	4	17.50	4.00	70	4	4.41	11.34	50
5	9.78	5.62	55	5	54.90	1.28	70	5	25.60	1.95	50
6	2.61	21.07	55	6	9.43	7.42	70	6	3.69	13.55	50
7	7.42	7.41	55	7	5.37	13.04	70	7	6.68	7.49	50
8	25.40	2.17	55	8	20.80	3.37	70	8	69.90	0.72	50
9	14.70	3.74	55	9	13.70	5.11	70	9	28.40	1.76	50
10	16.40	3.35	55	10	20.20	3.47	70	10	31.70	1.57	50
11	19.00	2.89	55	11	15.20	4.61	70	11	23.80	2.10	50
		79.93	605			77.42	770			74.07	550
Pool 7				Pool 8				Pool 9			
Amplicon	ng/µl	DNA vol	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)
1	5.47	10.60	58	1	6.69	17.94	120	1	14.40	5.21	75
2	32.30	1.80	58	2	43.40	2.76	120	2	50.80	1.48	75
3	4.99	11.62	58	3	6.26	19.17	120	3	4.41	17.01	75
4	5.68	10.21	58	4	11.80	10.17	120	4	11.40	6.58	75
5	10.90	5.32	58	5	8.00	15.00	120	5	11.70	6.41	75
6	4.66	12.45	58	6	8.42	14.25	120	6	61.20	1.23	75
7	2.79	20.79	58	7	9.32	12.88	120	7	3.81	19.69	75
8	19.40	2.99	58	8	38.40	3.13	120	8	30.80	2.44	75
9	34.40	1.69	58	9	31.10	3.86	120	9	34.00	2.21	75
10	37.90	1.53	58	10	25.60	4.69	120	10	48.10	1.56	75
11	11.60	5.00	58	11	21.40	5.61	120	11	26.40	2.84	75
		84.00	638			109.46	1320			66.66	825
Pool 10				Pool 11				Pool 12			
Amplicon	ng/µl	DNA vol	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)	Amplicon	ng/µl	DNA vol µl	Yield (ng)
1	11.20	17.86	200	1	12.70	5.91	75	1	10.70	7.94	85
2	44.30	4.51	200	2	40.60	1.85	75	2	31.60	2.69	85
3	11.20	17.86	200	3	20.20	3.71	75	3	4.78	17.78	85
4	12.70	15.75	200	4	5.34	14.04	75	4	8.18	10.39	85
5	18.60	10.75	200	5	4.56	16.45	75	5	39.20	2.17	85
6	11.30	17.70	200	6	7.86	9.54	75	6	4.11	20.68	85
7	15.20	13.16	200	7	3.67	20.44	75	7	19.70	4.31	85
8	37.30	5.36	200	8	33.10	2.27	75	8	27.40	3.10	85
9	33.20	6.02	200	9	32.30	2.32	75	9	26.60	3.20	85
10	37.90	5.28	200	10	38.90	1.93	75	10	14.10	6.03	85
11	17.50	11.43	200	11	22.00	3.41	75	11	18.40	4.62	85
		125.68	2200			81.87	825			82.91	935

Table 2.5 Normalisation and pooling of PCR Products. *PCR products are normalised and pooled in equimolar quantities each pool includes 11 BRCA1 amplicons. Each row represents one amplicon of BRCA1 Amplicon 1 to 11.*

PCR products are pooled in normalised (equalised) quantities to minimise biases in sequencing. This is done, as PCR reactions may not amplify template DNA with equal efficiency. Table 2.5 details how this is achieved.

2.14.2 Library validation

Following library preparation the libraries are validated using Agilent Bioanalyzer 2100. This measures DNA concentration and size of fragments in base pairs (bp). Figure 2.18 below is the output from the Agilent Bioanalyzer 2100, in these 12 samples the middle peak represents the DNA fragment size (in bp) and the concentration of DNA in each sample in nmol/l. The other two peaks represent DNA markers. Each of the samples are quantified in triplicate and the mean is used to calculate the DNA concentration required for pooling of 12 samples.

Figure 2.18 Library validation by Agilent Bioanalyzer 2100

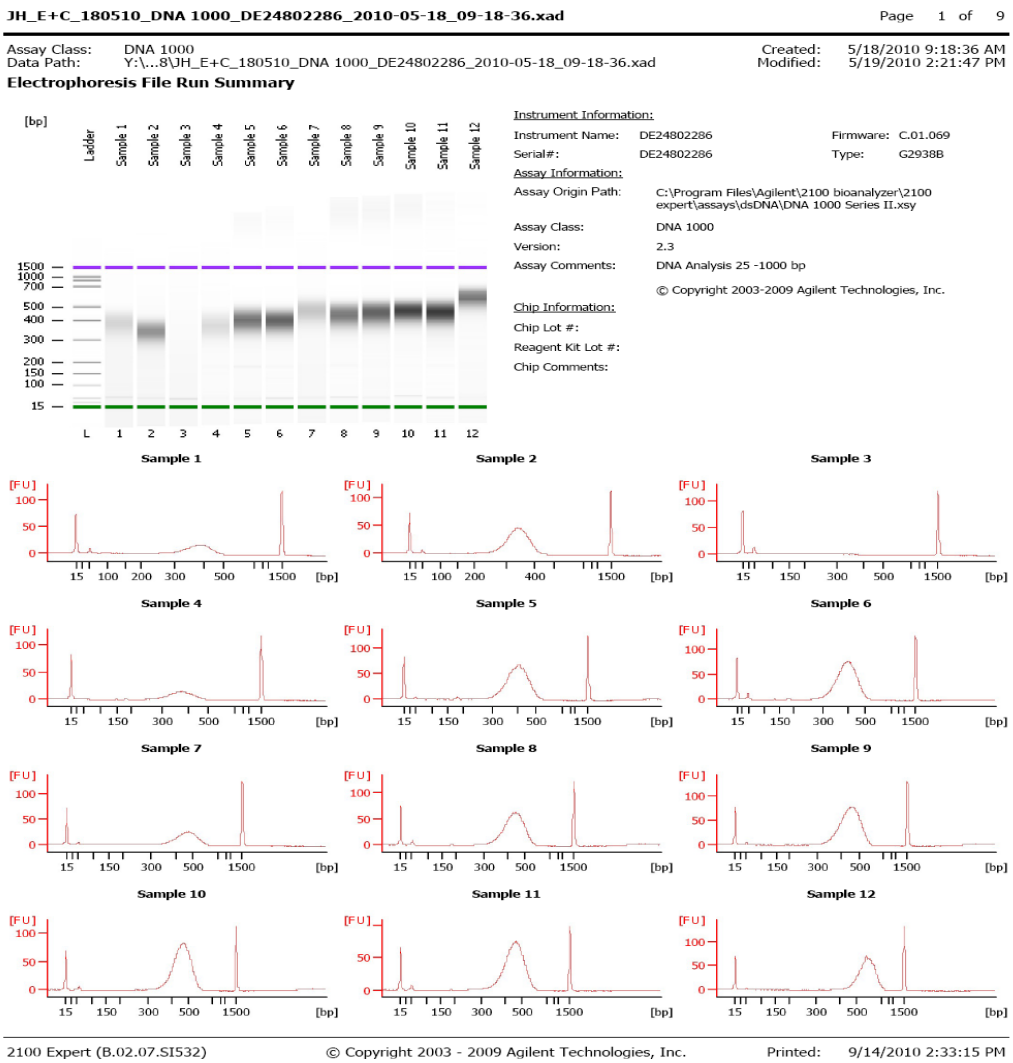


Figure 2.18 Library validation by Agilent Bioanalyzer 2100.

11 out of 12 libraries are validated on the Agilent Bioanalyzer 2100. Sample 3 is unfortunately not validated and this sample failed the library preparation. This is shown in Figure 2.18 where there is no peak at 300-400bp size band. The full Agilent trace output for triplicate repeats is in Appendix I. Quantitation and sizing is performed in triplicate and mean calculations used for normalisation.

2.14.3 Library normalisation

Table 2.6 Normalisation sheet for pooling 11 prepared libraries

Pool ID	Internal ID - External ID	Size	nM	nM	nM	Average nM	Pool Factor	DNA	Pool Volume	Index	Guideline Final concentration of Pool	Actual Final concentration of Pool
1	SOL567-1	390	18.5	21.7		20.10	78	3.88	23.08	1	50.03	58.13
	SOL568-2	344	46.5	45.5	44.2	45.40	78	1.72		2		
	SOL570-4	372	16.4	19.9		18.15	78	4.30		4		
	SOL571-5	404	64.5	68		66.25	78	1.18		5		
	SOL572-6	399	68.9	73.6		71.25	78	1.09		6		
	SOL573-7	478	19.7	17.5		18.60	78	4.19		7		
	SOL574-8	435	53.3	57.9		55.60	78	1.40		8		
	SOL575-9	464	59.6	70.8		65.20	78	1.20		9		
	SOL576-10	471	65.3	84		74.65	78	1.04		10		
	SOL577-11	464	76.1	78.7		77.40	78	1.01		11		
	SOL578-12	580	31	44.5		37.75	78	2.07		12		

Table 2.6 Normalisation sheet for pooling 11 prepared samples.

The results of the library validation are used to calculate the DNA input required to normalise samples for pooling. Table 2.6 describes the DNA quantities to create POOL1 of 11 patient DNA samples in equivalent quantities.

2.14.4 Agilent results for pool (concentration)

Following normalisation, the DNA concentration and size of fragments are analysed using the Agilent Bioanalyzer 2100 (Figure 2.19). This is performed in triplicate. This quality control shows that POOL1 has a mean molarity of 58.1 nmol/l at a mean size of 422 bp. This analysis is required to calculate DNA input for flow cell generation.

Figure 2.19 Quantitation of DNA concentration and sizing of pooled fragments

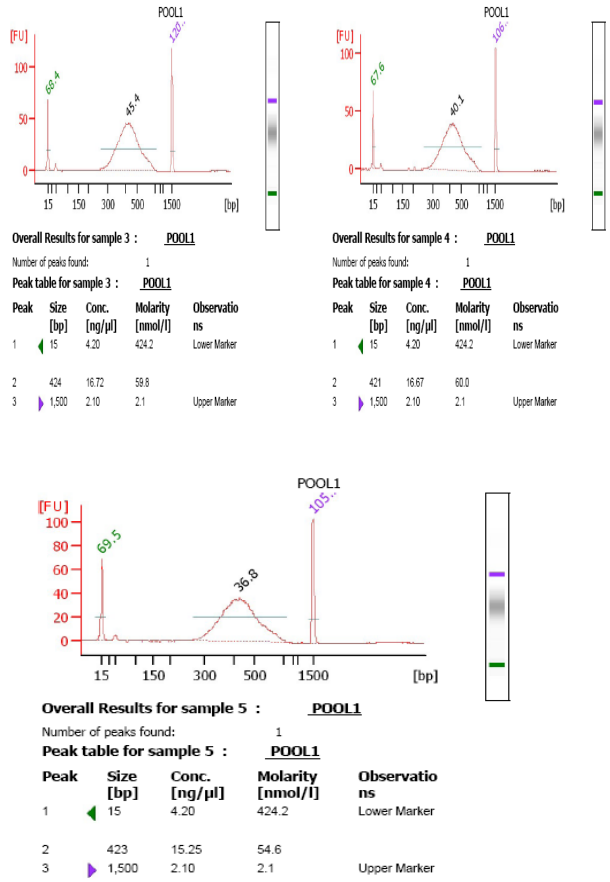


Figure 2.19 Quantitation of DNA concentration and sizing of pooled fragments.

2.14.5 Flow cell worksheet

Table 2.7 Library dilutions and DNA input for Pool

External and Dilutions for Cluster Station

Internal ID	POOL1	SOL541	SOL542	SOL543	SOL544	SOL511	SOL512	Phix
External ID	Index 1-12	s1	s2	s3	s4	2	3	
Ex conc.								
Ex 260/280								

Stock DNA (Library Prepared)

Internal ID	POOL1	SOL541	SOL542	SOL543	SOL544	SOL511	SOL512	Phix
Concentration								
260/280								
Agilent Size bp	423	388	389	366	364			
nM DNA	58.1	59.6	104.0	22.0	27.7	66.4	76.8	

10nM Stock Creation

Volume Required	60.0	50.0	100.0	20.0	30.0	20.0	20.0	
DNA	10.32	8.39	9.61	9.08	10.83	3.01	2.60	0.00
dH20	49.68	41.61	90.39	10.92	19.17	16.99	17.40	0.00

Table 2.7 Library dilutions and DNA input for pool

Table 2.8 Flow cell Generation

Flow Cell Generation 05/07/2010

	Internal ID	POOL1	SOL541	SOL542	SOL543	SOL544	SOL511	SOL512	Phix
	External ID	Index 1-12	s1	s2	s3	s4	2	3	0
	Final Concentration (pM)	12	12	12	12	12	12	12	4
	Adjustment Factor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	[Library] nM	10	10	10	10	10	10	10	
10nM Stock	NaOH 2N (µl)	1	1	1	1	1	1	1	0.5
10nM Stock	EB (µl)	16.00	16.00	16.00	16.00	16.00	16.00	16.00	8.50
10nM Stock	Template DNA (µl)	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00

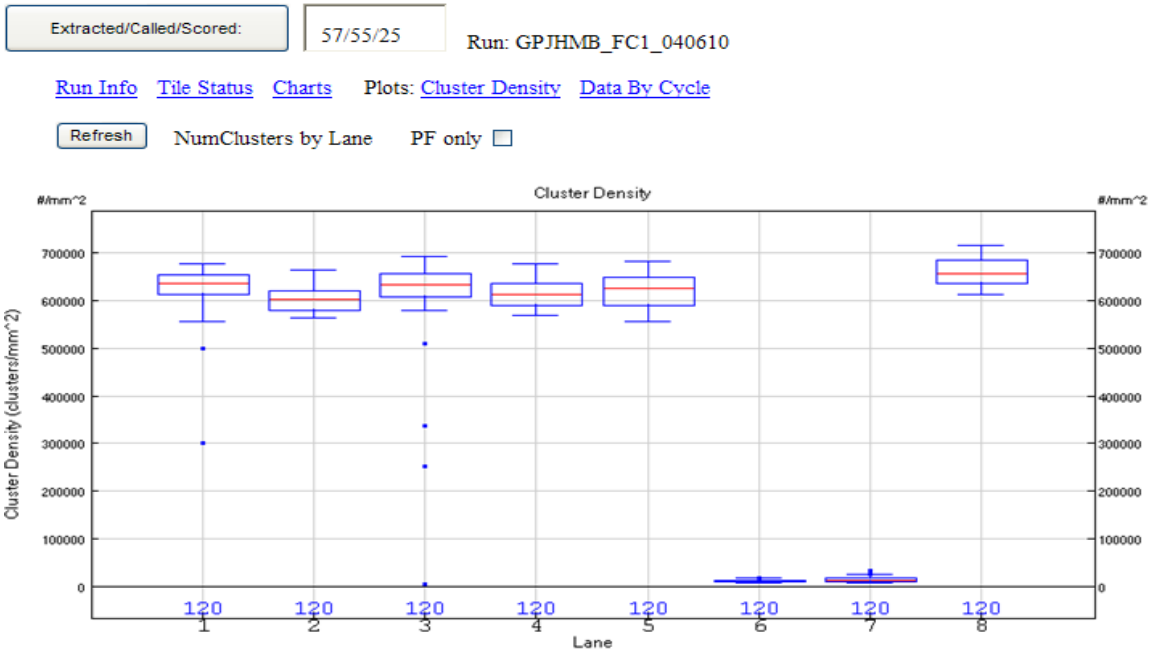
	FlowCell ID								
	Flow Cell Lane	1	2	3	4	5	6	7	8
Diluted Denatured	Denatured DNA	8	8	8	8	8	8	8	4
Diluted Denatured	Hybridisation Buffer	992	992	992	992	992	992	992	496
Diluted Denatured	Sequencing Primer								

Table 2.8 Flow cell Generation

The results from the Agilent Bioanalyzer 2100 are entered into a flow cell worksheet and final run volumes are calculated for DNA input. Note POOL1 is lane one of the flow cell. PhiX is lane 8 and is the control lane. The other six lanes are for another, unrelated, project.

2.14.5 Clustering results

Figure 2.20 Cluster density report



	Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8
MIN	556K	563K	579K	570K	557K	8K	8K	614K
MAX	678K	665K	693K	677K	682K	16K	24K	716K
P50	635K	602K	634K	614K	625K	12K	13K	658K

Figure 2.20 Cluster density report. *This cluster density report shows that the mean cluster density is 635,000 clusters per mm^2.*

In Figure 2.20 the first plot is the cluster density of POOL1. This shows that there is a minimum cluster density of 556,000 mm^2 and a maximum of 678,000 mm^2. To calculate the actual cluster density per tile, multiply this number by 0.53. So, 600,000 mm^2 x 0.53 = 318,000 clusters per tile and 38 million reads for lane one. There are 120 tiles per lane on Illumina GAIIx.

2.14.6 GAII Results

Table 2.9 The First base report

illumina®

Machine Name: 14-50-54
Run Date: 10-06-04
Run Id:

First Cycle (1)

MetricName	Lane1	Lane2	Lane3	Lane4	Lane5	Lane6	Lane7	Lane8
# of Clusters	248,367	245,973	254,893	250,845	254,870	2,743	2,427	265,648
Standard Dev	3,017	4,020	3,445	4,269	6,050	623	647	6,102
A Intensity	936	884	863	854	832	0	0	1,026
Standard Dev	7	12	17	55	46	0	0	206
C Intensity	838	785	765	750	731	401	277	891
Standard Dev	33	15	58	74	112	132	199	223
G Intensity	1,422	1,342	1,292	1,329	1,262	19	16	1,487
Standard Dev	53	27	82	107	94	19	16	270
T Intensity	1,592	1,502	1,468	1,480	1,410	8	0	1,634
Standard Dev	70	23	116	152	156	19	0	276
A Focus Metric	88	87	87	86	85	33	48	88
Standard Dev	1	1	1	2	4	19	29	4
C Focus Metric	81	80	81	79	78	73	72	81
Standard Dev	1	2	4	5	8	28	18	7
G Focus Metric	90	89	90	89	88	41	72	90
Standard Dev	1	1	2	2	4	23	21	4
T Focus Metric	90	88	89	88	87	39	50	90
Standard Dev	1	1	1	1	1	24	18	2
Foc Pos Min	-12,650	-11,470	-10,140	-9,140	-8,810	-7,980	-7,540	-7,530
Foc Pos Max	-5,330	-4,070	-2,910	-2,160	-1,850	-1,450	-1,190	-1,400
Flowcell Tilt	7,320	7,400	7,230	6,980	6,960	6,530	6,350	6,130

***Table 2.9 The first base report**. This report is produced as an initial quality control before continuing with the sequencing run.*

The Genome Analyser generates the first base report; this is produced so that quality control metrics can be assessed before continuing with the sequencing run. This report shows the number of clusters and the intensity of clusters for each of the 4 nucleotides.

2.15 Data analysis

2.15.1 Alignment of reads to human reference

Read alignment is performed using software Consensus Assessment of Sequence and Variation (CASAVA). This program is part of the Illumina GA data analysis workflow and converts the raw image data from the GA into aligned reads with base calls and quality control metrics.

2.15.2 Detection of variants using SAMtools (Sequence Alignment Map)

Table 2.10 Blinded causative mutations detected in NGS data.

Sample	Mutation	Exon	Mutation Type	HGVS Nomenclature
001. OVO25.515c	L78833.1:g37067del AA	11	Frameshift	3984delAA
002. Pr_B9	L78833.1:g35269delA	11	Frameshift	2187delA
003. Pr_B10	L78833.1:g36529delAAGC	11	Frameshift	3447del4
004. Pr_B1	Missing mutation			
005. OVO69.301b	L78833.1:g35079delAGTC	11	Frameshift	1997del4
006. OVO401.307	L78833.1:g34044delCTCA	11	Frameshift	962del4
007. Pr_B2	L78833.1:g24598G>T	7	Nonsense	E143X
008. OVO133.301b	L78833.1:g35854delTC	11	Frameshift	2773delTC
009. OVO89.305b	L78833.1:g34130delA	11	Frameshift	1048delA
010. OVO01.306b	L78833.1:g35154delA	11	Frameshift	2071delA
011. Pr_B7	L78833.1:g71668insC	20	Frameshift	5382insC
012. OVO34.411a	Library failed			

Table 2.10 Blinded causative mutations detected using NGS. *One library failed and 1 mutation is missed in one sample.*

On the second analysis using CASAVA and SAMtools the parameters are altered to include reads that are in regions of lower coverage (i.e. < 30 X). Table 2.10 shows the variants detected in the coding regions of *BRCA1* following the second analysis.

Table 2.11 All coding variants as identified by CASAVA software

Sample	Mutation	Exon	Amino Acid Change	Mutation Type	BIC Nomenclature
001. OVO25.515c	**L78833.1:g37067del AA**	**11**		**FS**	**3984delAA**
	L78833.1:g34953T>G	11			1871T>G
	L78833.1:g35589A>G	11			2507A>G
	L78833.1:g35801A>G	11			2719A>G
	L78833.1:g36320G>A	11	Ser1040Asn	MS	3238G>A
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
	L78833.1:g36429A>T	11			3347A>T
	L78833.1:g36905A>G	11			3822A>G
002. Pr_B9	**L78833.1:g35269delA**	11		**FS**	**2187delA**
	L78833.1:g35283C>T	11			2201C>T
	L78833.1:g35512T>C	11			2430T>C
	L78833.1:g35813C>T	11	Pro871Leu	MS	2731C>T
	L78833.1:g36314A>G	11	Glu1038Gly	MS	3232A>G
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
	L78833.1:g36749A>G	11	Lys1183Arg	MS	3667A>G
	L78833.1:g46278T>C	13			4427T>C
	L78833.1:g57655A>G	16	Ser1613Gly	MS	4956A>G
003. Pr_B10	L78833.1:g35283C>T	11			2201C>T
	L78833.1:g35512T>C	11			2403T>C
	L78833.1:g35813C>T	11	Pro871Leu	MS	2731C>T
	L78833.1:g36314A>G	11	Glu1038Gly	MS	3232A>G
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
	L78833.1:g36529delAAGC	**11**	Gln1069Leu	**FS**	**3447del4**
	L78833.1:g36749A>G	11	**Lys1183Arg**	MS	3667A>G
	L78833.1:g46278T>C	13			4427T>C
	L78833.1:g57655A>G	16	Ser1613Gly	MS	4956A>G
004. Pr_B1	Missing mutation				
	L78833.1:g4708T>G	2	Cys24Gly	MS	189T>G
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
005. OVO69.301b	L78833.1:g34268A>G	11	Gln365Arg	MS	1186A>G
	L78833.1:g35079delAGTC	**11**		**FS**	**1997del4**
	L78833.1:g35283C>T	11			2201T>C
	L78833.1:g35512T>C	11			2430T>C
	L78833.1:g35813C>T	11	Pro871Leu	MS	2731C>T
	L78833.1:g36314A>G	11	Glu1038Gly	MS	3232A>G
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
006. OVO401.307	**L78833.1:g34044delCTCA**	**11**		**FS**	**962del4**
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T

Sample	Mutation	Exon	Amino Acid Change	Mutation Type	BIC Nomenclature
007. Pr_B2	L78833.1:g24598G>T	7	Glu to Stop	NS	E143X
	L78833.1:g36388T>C	**11**	**Ser to Pro**	**MS**	**3306T>C**
	L78833.1:g36389C>A	**11**	**Ser to Tyr**	**MS**	**3307C>A**
	L78833.1:g36391A>T	**11**	**Ser to Cys**	**MS**	**3309A>T**
	L78833.1:g36396T>A	**11**	**Asp1065Glu**	**MS**	**3314T>A**
	L78833.1:g36404T>A	**11**	**Ile1068Asn**	**MS**	**3322T>A**
	L78833.1:g36405T>G	**11**	**Ile1068Met**	**MS**	**3323T>G**
	L78833.1:g36407A>T	**11**	**Gln1069Leu**	**MS**	**3325A>T**
008. OVO133.301b	**L78833.1:g35854delTC**	11		**FS**	**2773delTC**
	L78833.1:g36396T>A	11	**Asp1065Glu**	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
009. OVO89.305b	**L78833.1:g34130delA**	**11**		**FS**	**1048delA**
	L78833.1:g35283C>T	11			2201C>T
	L78833.1:g35512T>C	11			2403T>C
	L78833.1:g35813C>T	11	Pro871Leu	MS	2731C>T
	L78833.1:g36314A>G	11	Glu1038Gly	MS	3232A>G
	L78833.1:g36396T>A	11	Asp1065Glu	MS	3314T>A
	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
	L78833.1:g36749A>G	11	Lys1183Arg	MS	3667A>G
	L78833.1:g46278T>C	13			4427T>C
	L78833.1:g57655A>G	16	Ser1613Gly	MS	4956A>G
010. OVO01.306b	**L78833.1:g35154delA**	**11**		**FS**	**2071delA**
	L78833.1:g35283C>T	11			2201C>T
	L78833.1:g35512T>C	11			2403T>C
	L78833.1:g35813C>T	11	Pro871Leu	MS	2731C>T
	L78833.1:g36314A>G	11	Glu1038Gly	MS	3232A>G
	L78833.1:g36749A>G	11	Lys1183Arg	MS	3667A>G
	L78833.1:g46169T>C	13			4318T>C
	L78833.1:g46173A>G	13			4322A>G
	L78833.1:g46278T>C	13			4427T>C
	L78833.1:g57655A>G	16			4956A>G
011. Pr_B7	L78833.1:g36404T>A	11	Ile1068Asn	MS	3322T>A
	L78833.1:g36405T>G	11	Ile1068Met	MS	3323T>G
	L78833.1:g36407A>T	11	Gln1069Leu	MS	3325A>T
	L78833.1:g71668insC	**20**		**FS**	5382insC
012. OVO34.411a	*Library failed*				

Table 2.11 All coding variants as identified by CASAVA software*. Likely sequencing artefacts are highlighted in red. MS = missense, FS = Frameshift, NS = Nonsense*

The Table 2.11 shows a list of all coding variants that are detected by NGS. Many of these are likely to be sequencing artefacts and these are evident as the parameters are changed to include all variants in regions that are below 30X depth of coverage. These are considered artefacts as either they are in sequence next to each other i.e. each nucleotide in a sequence is called as single base change or they are repeated in many samples. This suggests that the alignment is incorrect. Interestingly, where there is a missing mutation in exon 2 of sample Pr_B1, CASAVA has called a single base change in that position. It is likely that this is part of the missing mutation and that it is missed due to incorrect alignment during read mapping. In addition, no minimum percentage for alternate allele is set. These issues are addressed in the next study.

2.15.3 Detection of unclassified variants in regulatory and intronic regions

Many unclassified variants are detected, including insertions and deletions in the promoter region and 3' UTR, as well as in introns. These large data sets are available as excel files; however, the files are too large to print as each sample has ~300-400 novel variants in non-coding regions. Many of these are single base changes and are repeated in several samples. The Table 2.12 describes unclassified intronic variants, which are insertions or deletions of bases 2-5 bp in size that are unique to samples.

Table 2.12 Unclassified intronic variants

Sample	Intronic variant	Intron
001. OVO25.515c	22806insCTTG 42620delTA	5 12
002. Pr_B9	9446delTT 32346insTTTC 42616insTGTT 60880insAAC 74964delTTC	2 9 12 16 20
005. OVO69.301b	69249insGTA	19
007. Pr_B2	5359delCA	2
008. OVO133.301b	37797delGT	12
009. OVO89.305b	4827insGGG 79985insGG 82163delTAA	13 22 23
010. OVO01.306b	21495insAC 22806insCTCA 4827insTGGG 60603insTT 73746delCC	3 5 13 16 20
011. Pr_B7	27768delCT 28683delCC 54064del AT 69250insTTT	7 7 15 19

Table 2.12 Unclassified intronic variants. *Variants selected are those that are 2-5 bp indels. The table shows the nucleotides inserted or deleted and the location of the variant.*

The Table 2.12 above details some examples of changes found in intronic regions of *BRCA1*. The significance of these is difficult to determine here and their relevance in cancer susceptibility is uncertain unless RNA testing is conducted to determine if these intronic changes result in a shorted RNA transcript. In addition, many of these may be artefacts of the sequencing technology.

2.15.4 Capillary sequencing of exon 2 in sample Pr_B1

Figure 2.21 Capillary sequencing trace of the end of exon 2

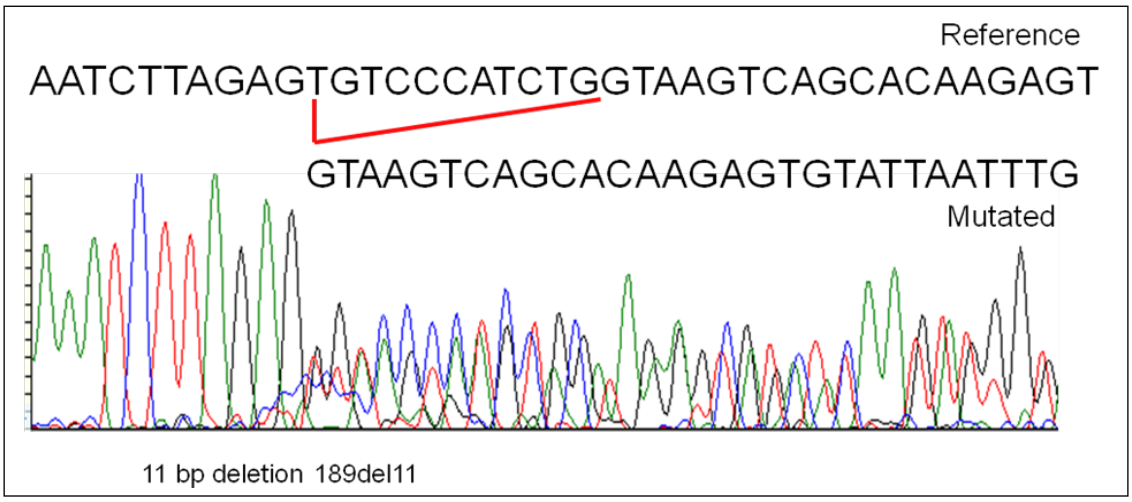


Figure 2.21 Capillary sequencing trace of the end of exon 2. *This trace shows the deletion of the last 11bp of exon 2 in sample Pr_B1*

One mutation is not detected by NGS. To troubleshoot this missing mutation capillary sequencing of the region is performed to verify the mutation in the sample, which is confirmed. Figure 2.21 shows the mutation 189del11 in sample Pr_B1.

2.15.5 A revised analysis pipeline for the re-analysis of Exon 2 in sample Pr_B1

The overwhelming growth in the sequencing capacity of massively parallel sequencing technologies has resulted in the necessary development of new computational methods to analyse resultant large sequencing data sets. These tools are continually being refined and developed at an astounding rate. Many of these tools are specifically designed for certain tasks, i.e. one for read mapping, another for SNP detection and another for indel detection. One difficulty with this is that they may not be compatible on all platforms. Sequencing capacity and data analysis are not developing at an equal rate, meaning that data analysis methods are not easily extracting the required information to answer specific biological questions. Downstream analysis requires the manipulation of sequencing data in order to extract answers to these complex questions. It follows that in any analysis pipeline there must be adaptability that allows for programming to be performed thus, tailoring the analysis to the individual experiment (McKenna et al 2010).

Almost certainly the most crucial step in analysing massively parallel sequencing data is the first one that involves mapping millions of short 100bp reads individually to the reference sequence. Immediately, it is obvious that if reads contain an indel they run

the risk of being incorrectly mapped. Although each base is assigned a quality score that calculates the likelihood that the base is the correct sequenced base, this cannot be relied upon. These quality scores are renowned for their imprecision (DePristo et al 2011). Both of these major flaws can have serious consequences and produce inaccuracies in variant detection. The main obstacle to overcome in the analysis of NGS data is discerning genuine genetic variation from artefacts of the sequencing technology.

To resolve these issues analysis tools are available to filter reads on the basis of certain parameters. For example, variants detected in regions of too high or too low coverage can be filtered out. In this study, it is shown that both of these pose a problem. In the regions of very high coverage due to excess sequencing capacity and the use of Long Range PCR, filters are changed to allow for the detection of several variants.

A new analysis pipeline is used for the investigation of the missing 189del11 mutation in sample Pr_B1. This revised analysis pipeline is performed on sample Pr_B1. The analysis pipeline of the pilot study and the revised pipeline are described in the following flow diagrams:

Figure 2.22 Original analysis pipeline: pilot study

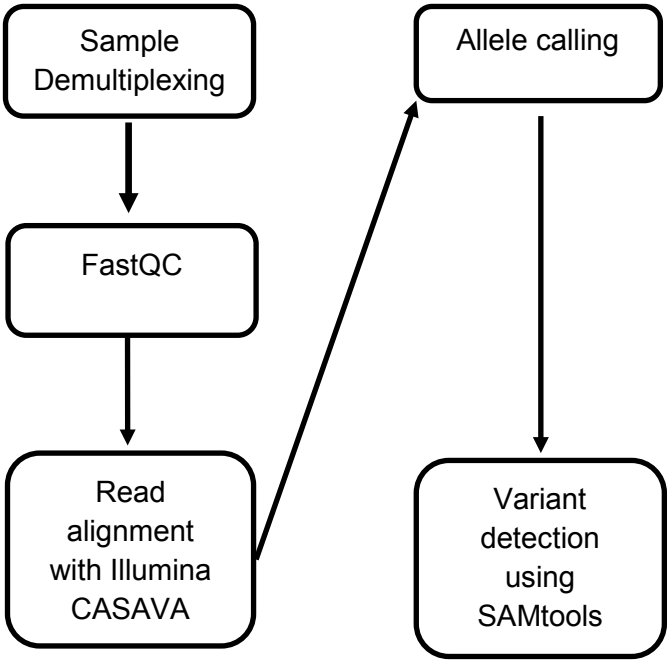


Figure 2.22 The original analysis pipeline. *This format is used in first analysis.*

Figure 2.23 Revised analysis pipeline

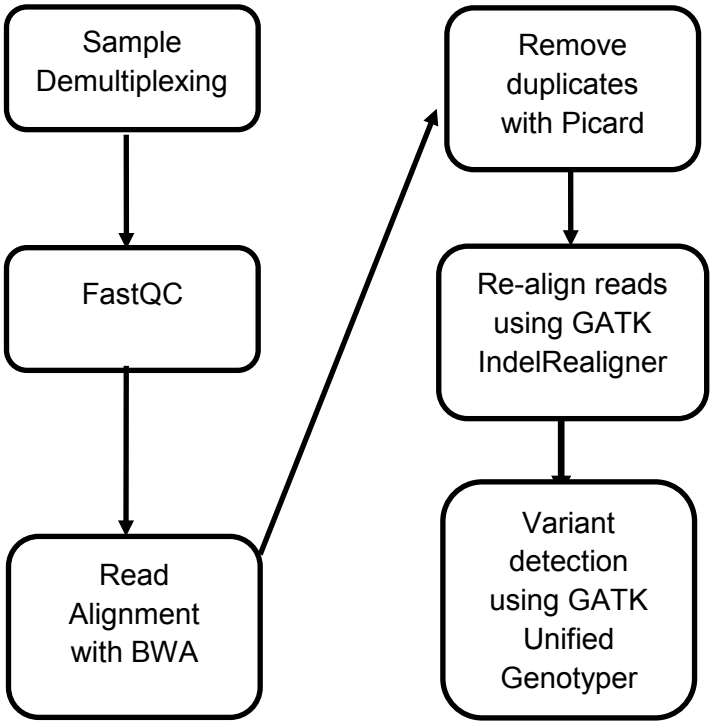


Figure 2.23 The revised analysis pipeline*. This is the format used in the re-analysis of exon 2 sample Pr_B1.*

2.15.6 Re-analysis of Pr_B1 exon 2 with an alternative analysis pipeline

2.15.6.1 FastQC

FastQC performs quality control (QC) analysis on raw data generated from high-throughput sequencing. This program effectively highlights any areas of concern in the data. This Java based program analyses the raw data and graphically reports a detailed image of sequence quality specifically for high-throughput sequencing data. QC data are displayed and the following describes the information provided by the program:

1. Basic Statistics

These statistics (Table 2.13) give brief information on sequence ID, library sequence length, number of sequences and GC% content of sequences.

2. Per base sequence quality

This displays quality scores per base (Figure 2.24); scores should be high and at least over 20. Yellow blocks show the 25^{th} to 75^{th} percentile and the blue line is the mean.

3. Per sequence quality scores

This shows the mean across all bases and plots the distribution of means (Figure 2.25)

4. Per base sequence content

This graph shows per base sequence content and will demonstrate if the position of a base in the sequence influences base calls. Lines should be roughly parallel, as position should not influence base call (Figure 2.26).

5. Per base GC content

This line should be flat, as % GC content should not vary through sequence reads (Figure 2.27).

6. Per sequence GC content

This graph should show a normal distribution (Figure 2.28). Peaks could demonstrate contaminants.

7. Per base N content

This should be a flat line at 0% to show that no base calls were N, meaning that a base is not identified (Figure 2.29).

8. Sequence length distribution

This plot demonstrates if all sequences are the same length i.e. 76 bp (Figure 2.30)

9. Sequence duplication levels

This plot shows how unique sequences are (Figure 2.31). A duplication level of 1 shows that it is mostly unique and a duplication level of 10 is 10 x duplicated and so is not unique.

10. Overrepresented sequences

This table (Table 2.14) shows a list of those overrepresented sequences with the percentage of each sequence making up the library.

2.15.6.2 Burrows Wheeler Aligner (BWA)

The Burrows-Wheeler Aligner (BWA) is a read-mapping program that effectively aligns short reads (up to 200 bp) against a reference sequence (Li & Durbin 2009). It allows for gapped alignment, which is crucial when aligning single read sequences that may include insertions and deletions. BWA is used in a third analysis of just one sample Pr_B1 in order to troubleshoot the missing mutation in exon 2 that is not identified in previous analyses.

2.15.6.3 SAM (Sequence Alignment Map) Format and SAMtools

The SAM format is a universal format that stores NGS reads that are aligned using read mapping programs. These various read mapping programs produce read alignment data in many different, and incompatible, formats. Converting aligned reads to SAM format allows for the further downstream analysis of sequencing data to effectively detect gene variation. These downstream analyses can be performed using SAMtools and Picard. SAM has a corresponding binary format known as BAM, which enables compression of data.

2.15.6.4 Picard

Picard is a tool that can manipulate files in SAM format. Following alignment with BWA, Picard is used to remove the duplicate sequences. These duplicate sequences include the indexes used to identify individual samples. They also include the overrepresented sequences that appear to be duplicated at each end of the LR-PCR fragment. These sequences can be cleaned up (i.e. removed) by Picard and realigned using the Genome Analysis Toolkit developed by the Broad Institute.

2.15.6.5 Genome Analysis Toolkit (GATK) Broad Institute

The GATK is a programming structure designed to provide tools that enable the more refined analysis of next generation sequencing data using the programming model MapReduce (DePristo et al 2011). This software framework facilitates the writing of analysis tools for MPS resequencing data. GATK has been used for the analysis of the 1000 Genome Project data and the Cancer Genome Atlas.

2.15.6.6 Realignment with GATK IndelRealigner

Local realignment is often required around regions that include indels. The Multiple Sequence Alignment (MSA) tool realigns reads to reduce the number of mismatched bases across the read. If these mismatches are not removed they can be erroneously identified as SNPs. This is the case in this study as the CASAVA program identifies two single base changes in the region that the 11 bp deletion resides rather than identifying the mutation. This is due to the incorrect alignment of reads containing the deletion. Once local realignment is performed the resulting reads are clean and can be put through the variant detection program (Unified Genotyper, which is a tool available from GATK).

Local realignment is performed in two stages: first is to establish intervals that may require realignment. So, in this case exon 2 is identified because, the indel is expected be located in exon 2. Second, the realigner is run over the established interval (exon 2). In the re-analysis of exon 2 Pr_B1 100bp intervals are used and run through the realigner.

2.15.6.7 Variant detection using GATK Unified Genotyper

This tool calls SNPs and indels from both single sample and multi-sample data. It attempts to establish both the likely genotypes and the allele frequencies within the input population of samples.

2.15.6.8 Integrative Genomics Viewer (IGV)

The Integrative Genomics Viewer is a data visualisation tool, also from the Broad Institute, that is fully compatible with GATK. BAM or SAM files can be imported directly and visualised alongside RefSeq genes. This tool allows the user to interrogate data sets from genomics experiments, including sequencing, copy number variation and gene expression data (Robinson et al 2011)

2.15.7 Fast QC data

Table 2.13. Basic statistics

Measure	Value
Filename	s_1_sequence.txt
File type	Conventional base calls
Total Sequences	795819
Sequence length	76
%GC	44

Table 2.13 Basic statistics table. *This table shows that the sample file contains 795,819 sequences of 76 bp in length.*

Figure 2.24 Per base sequence quality

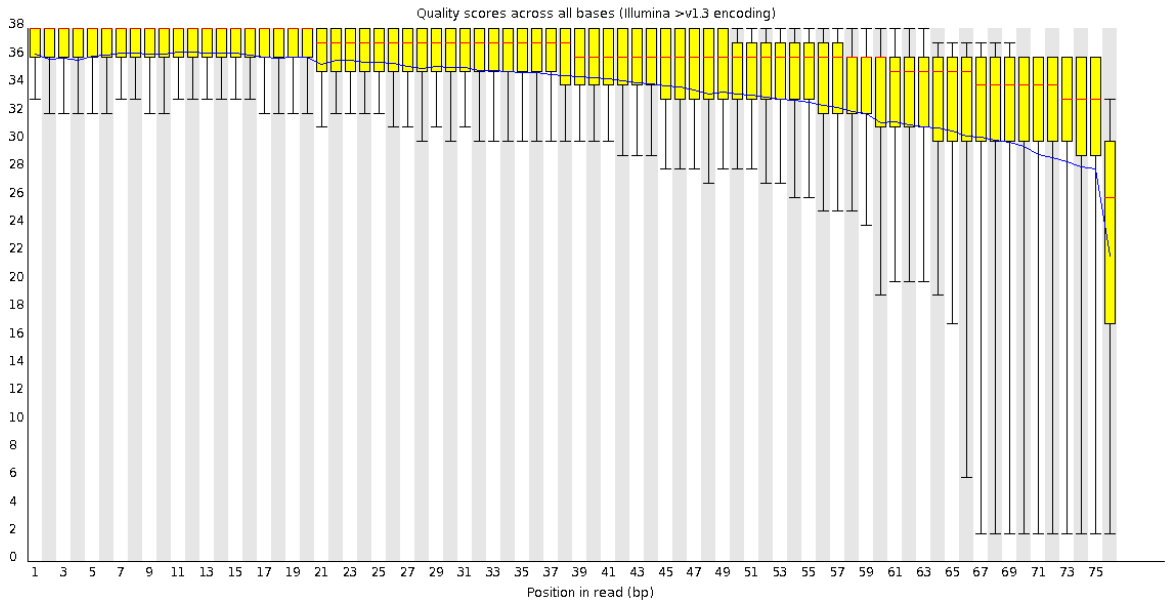


Figure 2.24 Per base sequence quality. *This image shows that mean quality scores are all high.*

Figure 2.25 Per sequence quality scores

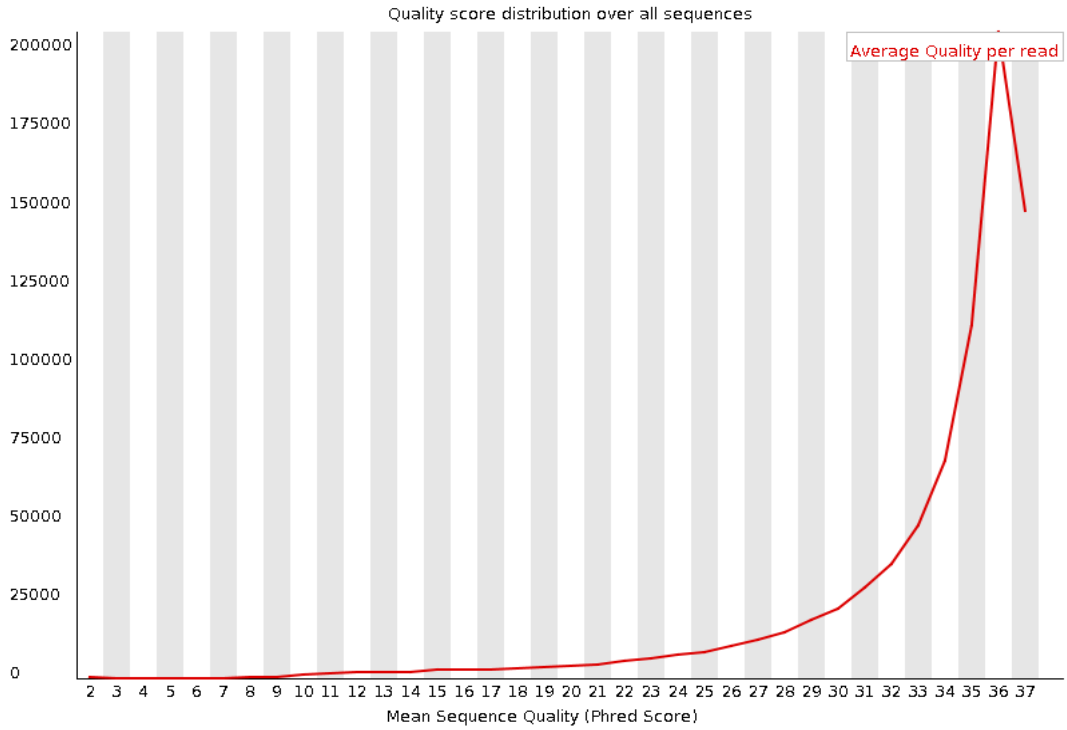


Figure 2.25 Per sequence quality scores. *This image shows that the mean quality scores are high*

Figure 2.26 Per base sequence content

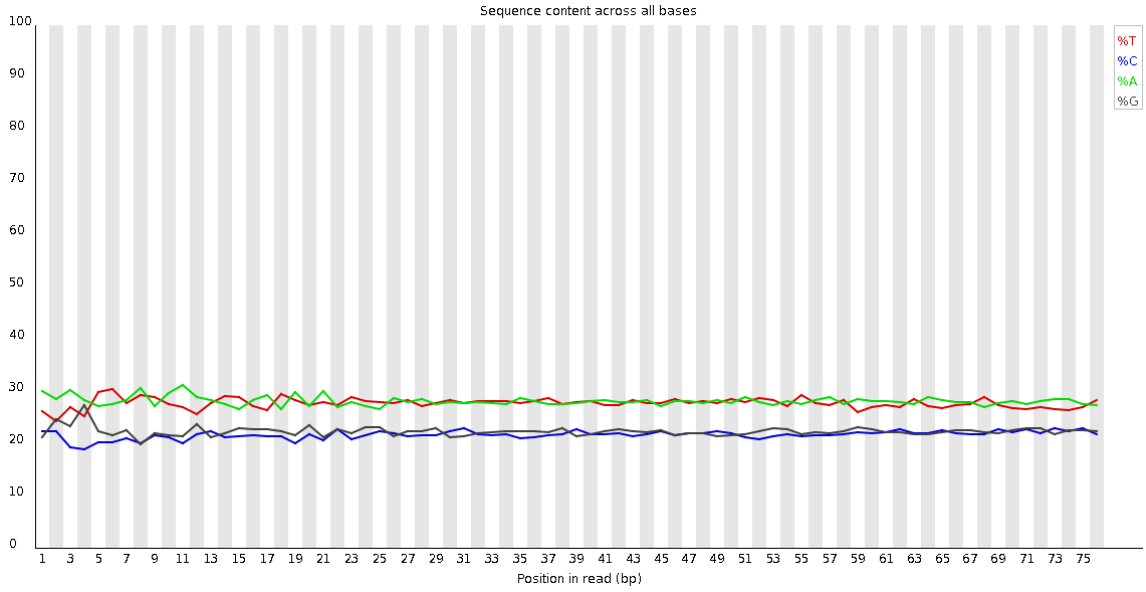


Figure 2.26 Per base sequence content. *This image shows that the position of the base does not influence base call. The four bases are equally called.*

Figure 2.27 Per base GC content

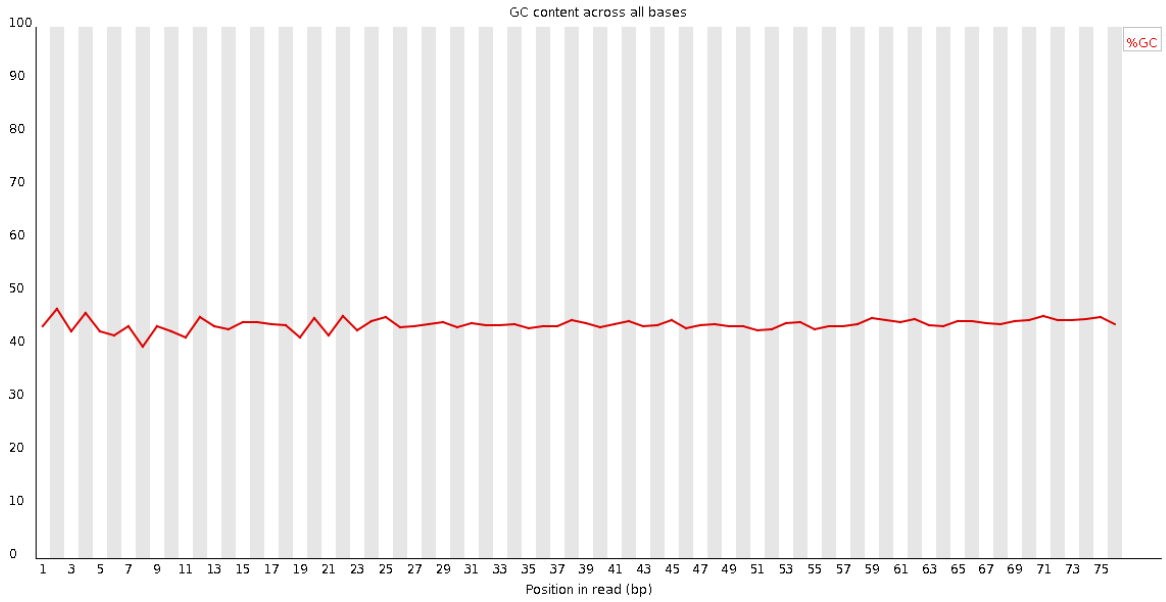


Figure 2.27 Per base GC content. *This image shows that the per base GC content does not vary between reads.*

Figure 2.28 Per sequence GC content

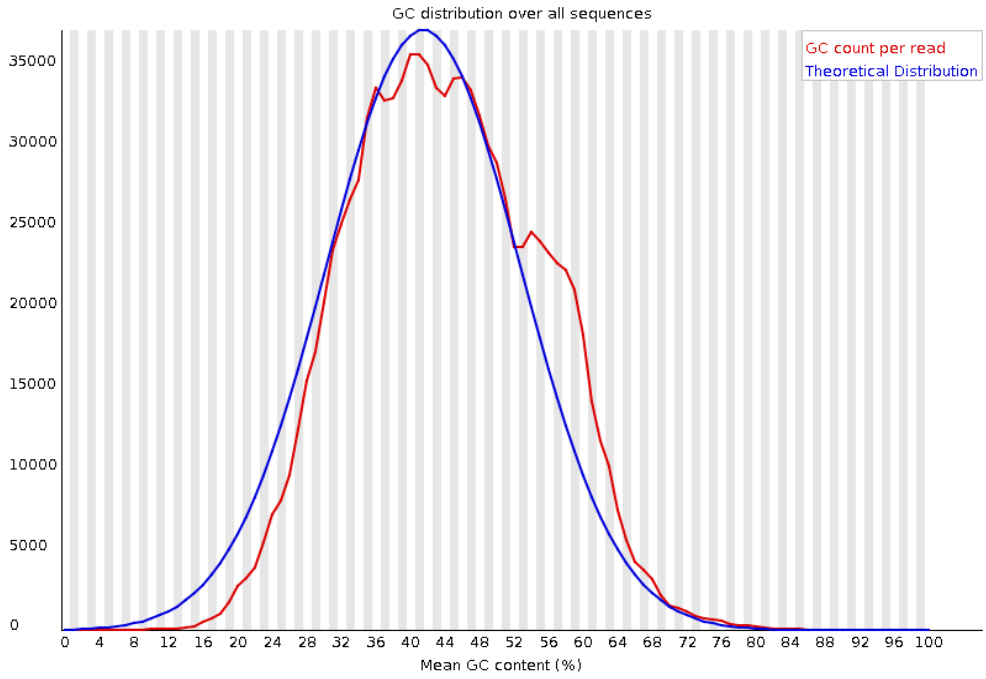


Figure 2.28 Per sequence GC content. *This image shows that the distribution is roughly a normal distribution. The small peak on the right indicates a level of contamination.*

Figure 2.29 Per Base N content

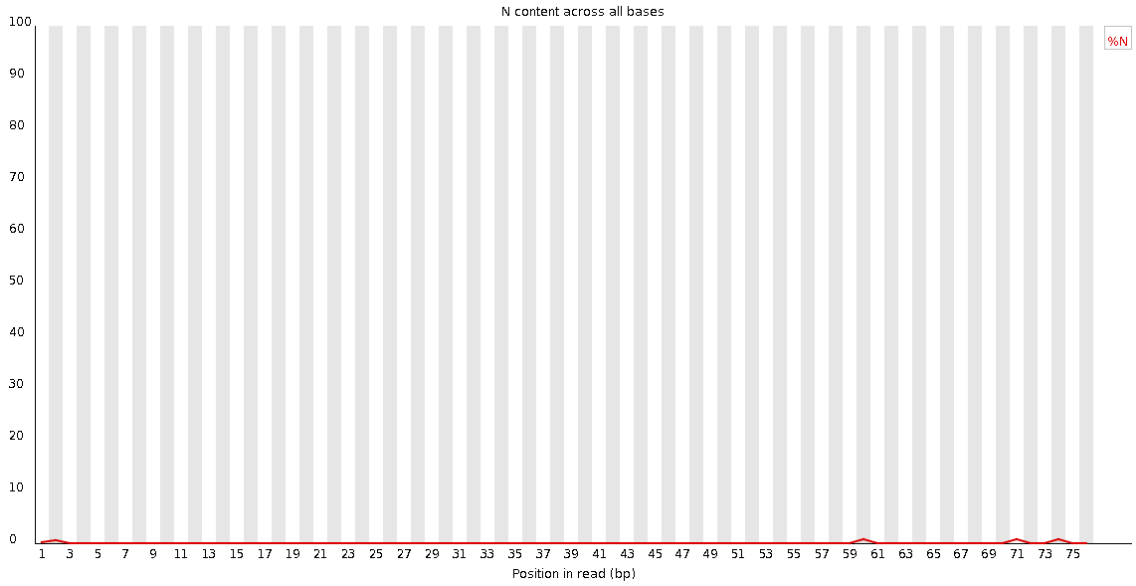


Figure 2.29 Per Base N content. *This image shows that the N content is minimal; most bases are called.*

Figure 2.30 Sequence length distribution

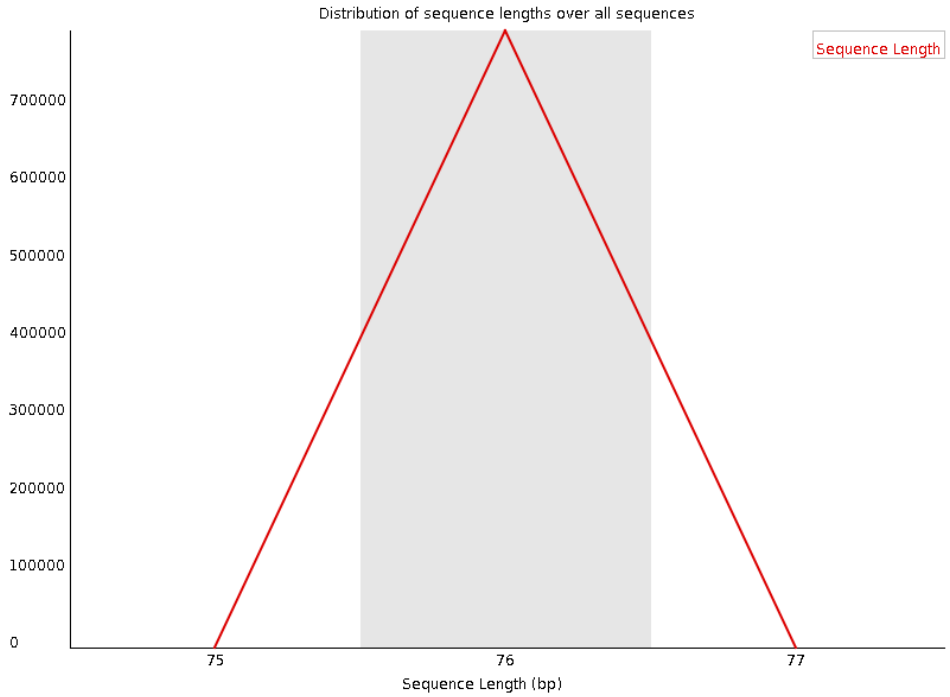


Figure 2.30 Sequence length distribution. *This image shows that sequences are all the same length*

Figure 2.31 Sequence duplication levels

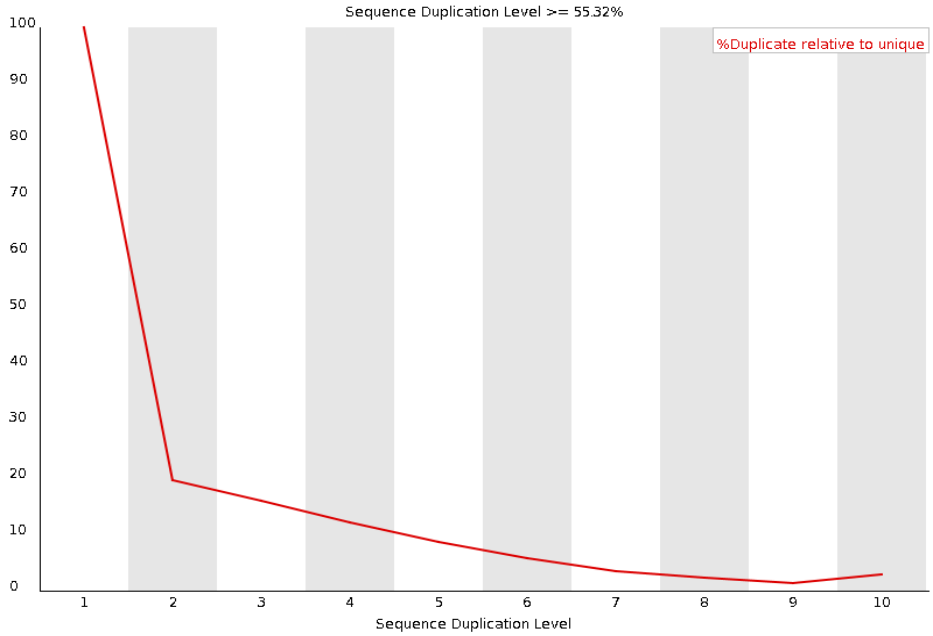


Figure 2.31 Sequence duplication levels. *This plot demonstrates that 55.32% of the sequences are not unique.*

Table 2.14 Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGCCTTGTCTCAGCTGGGTGTCTTTATTTACTCTGTCTTAAAGT GTTCCTTTTATTATCATTATTATTTTTTAATC	2328	0.29252882879147146	No Hit
GGCTTGTAAGAATGCCCTGCCACTTCTGCCCTGCAATATCCCT TGCTATTAGGATTTTGGCATCACCTTGGGTCCT	2177	0.2735546650683133	No Hit
ACAATTCAGAGCAGGGGTAGGGAGGAAATTCCCTTATGATAGT ACTGCAGAATATAGTACAGTAGAGTGACAAGCT	1871	0.2351037107684034	No Hit
GCAGGGCAGAAGTGGCAGGGCATTCTTACAAGCCAGGATGAA AACAAACACTAGAGAAATGCTACTATCTGGCAGT	1769	0.22228672600176674	No Hit
TGTATAGACTACAGCACGAGACAGCTTAGCTTGTCACTCTACT GTACTATATTCTGCAGTACTATCATAAGGGAAT	1767	0.22203541257496992	No Hit
CCAGACATTTTAGTGTGTAAATTCCTGGGCATTTTTTCCAGGCA TCATACATGTTAGCTGACTGATGATGGTCAAT	1733	0.2177630843194244	No Hit
ACAGCACTTGAGTTGCATTCTTGGGATATTCAACACTTACACTC CAAACCTGTGTCAAGCTGAAAAGCACAAATGA	1712	0.21512429333805802	No Hit
CCTAGTGCCCAGAACACAGTAGGCTCCCAATAAATAGCCACTG AATAAAAGTTAAAACCAACAAAAATAATCATTT	1573	0.19765801017568063	No Hit

CACTGTGAAGAAAACAAGCTAGCAGAACATTTTGTTTCCTCACT AAGGTGATGTTCCTGAGATGCCTTTGCCAATA	1438	0.18069435386689686	No Hit
AGAGGAGACAAGGAGCATGTACACCTAAAATCAACATAGACCC CTCTGTTGATGGGGTCATAGTGAGTACTTGAGG	1165	0.14639007110913413	No Hit
AGCCCTTTCACCCATACACATTTGGCTCAGGGTTACCGAAGAG GGGCCAAGAAATTAGAGTCCTCAGAAGAGAACT	1090	0.13696581760425422	No Hit
ACGACTAACCTGGCAGTGTGACAAGAATGTGGTTTTTTCCTTAA ATATTTAACTTTTTAGAAAAGGATCACAAGGG	1067	0.13407571319609107	No Hit
CAGGTTATGTTGCATGGTATCCCTCTGCTTCAAAAACGATAAAT GGCACCAAGAAAATGAAATACTTTGAGAAGCT	1052	0.1321908624951151	No Hit
CAGGTTGCTGGCCCCACCTGTCTGGGATTCAGTGGGTCTGGG AATTTGCATATCTAACAAATTCCTAGGTGATGGT	1004	0.126159340251992	No Hit
CCAGACATTTTAGTGTGTAAATTCCTGGGTTCTGCTAGCTTGTT TTCTTCACAGTGAGATCGGAAGAGCACACGTC	987	0.12402317612421919	No Hit
CAGGTTGCTGGCCCCACCTGTCTGGGCTGTCTCAACAGTTTTG GGTTTGCCTGGATTTCACTGCATCTTGAGCTGG	870	0.10932134065660659	No Hit
CACTGTGAAGAAAACAAGCTAGCAGAACCCAGGAATTTACACA CTAAAATGTCTGGAGATCGGAAGAGCACACGTC	836	0.10504901240106104	No Hit
AGTGTCACCACCCCAAGGACTCTCTCTCATTTTCTTTGCCTGG GCCCTCTTTCTACTGAGGAGTCGTGGCCTTCCA	831	0.10442072883406907	No Hit

***Table 2.14 Overrepresented sequences**. This table shows the sequences that are overrepresented.*

The sequences in Table 2.14 are those that are overrepresented. On closer inspection of these sequences they are the beginning of LR-PCR amplicons.

Following QC checks, the reads from sample Pr_B1 are aligned with BWA, and then duplicate sequences are removed with Picard tools. Next the cleaned up sequences are re-aligned with GATK IndelRealigner and variant detection is performed on exon 2 Pr_B1 using GATK. The Figure 2.32 overleaf shows the reads using the Integrative Genome Viewer (IGV). Here the 11 bp deletion at the 3' end of exon 2 is clearly identified. The second IGV image in Figure 2.33 shows the table zoomed out and demonstrates that there are 7 high quality reads that identify the 11bp deletion.

Figure 2.32 The Integrative Genomics Viewer (IGV)

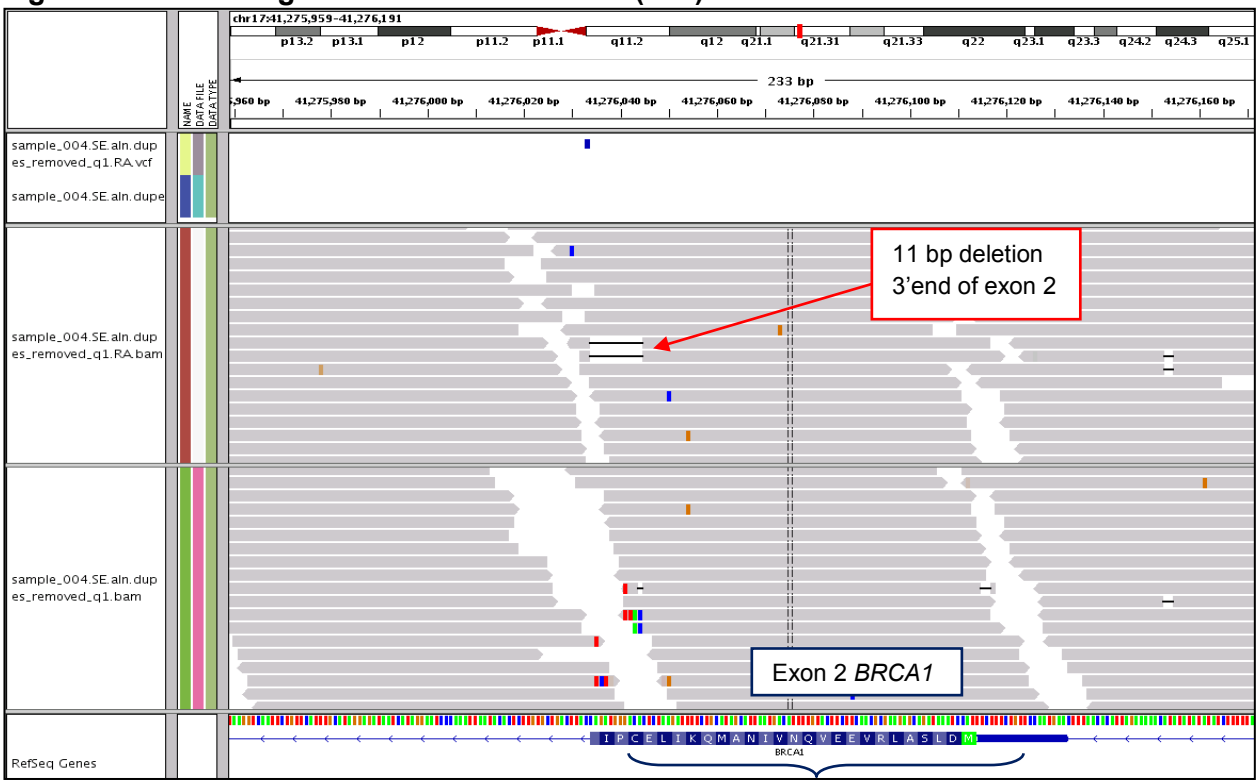


Figure 2.32 The Integrative Genomics Viewer (IGV). *This image shows the 11 bp deletion at the end of exon 2.*

Figure 2.33 The Integrative Genomics Viewer (IGV)

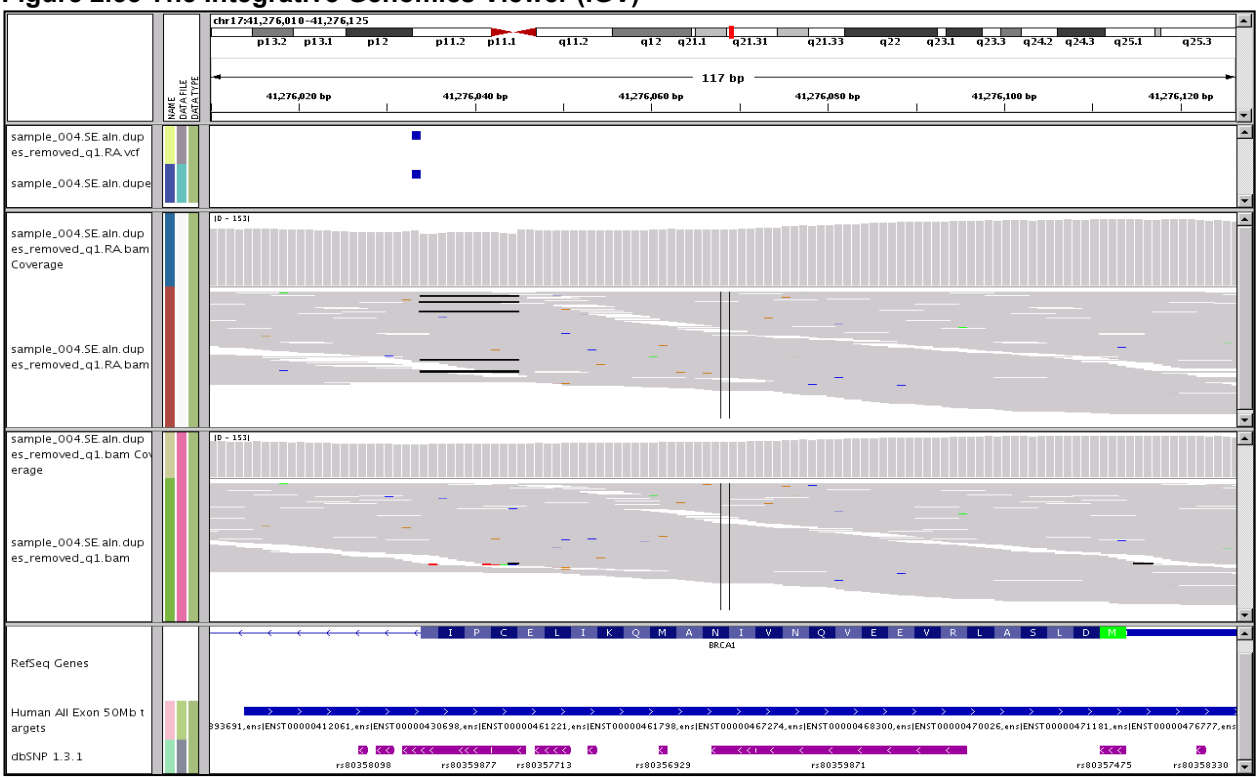


Figure 2.33 The Integrative Genomics Viewer (IGV). *This view of IGV is zoomed out, with clearly visible 7 high quality reads containing the 11 bp deletion.*

2.15.8 Troubleshooting the missing 11 bp deletion in Pr_B1

A third Analysis of Exon 2 of sample Pr_B1 is conducted to ascertain if the 11 bp deletion in exon 2 of Pr_B1 can be detected in the Illumina GAIIx sequencing data. This analysis is based on the alternative pipeline outlined in the Methodology chapter.

The Fast QC checks for Pr_B1 are mostly passed, however one highlighted warning is the number of repeated sequences (Table 2.14). Pr_B1 results show that 55.32% of the sequences are duplicated, meaning they were not unique sequences. On closer examination of the individual sequences the overrepresented sequences are the beginning of LR-PCR amplicons. Other researchers using LR-PCR and Illumina sequencing observe this; in that the 50 bp at each end of LR-PCR fragment can represent in excess of 50% of the sequenced bases (Harismendy & Frazer 2009). Reducing the overrepresentation of amplicon ends will improve coverage uniformity and result in an overall increase in sequencing depth across the sequencing target.

The new analysis pipeline cleans up duplicated sequences before realigning reads. This results in the detection of the 11bp indel in the sample. This demonstrates that by using a different analysis pipeline the missing mutation in the pilot study is in fact detected. This allows more confidence in the data analysis as this study detected 100% of mutations in samples sequenced using Illumina GAIIx.

2.15.9 Coverage data

Figure 2.34 Coverage data for one patient sample

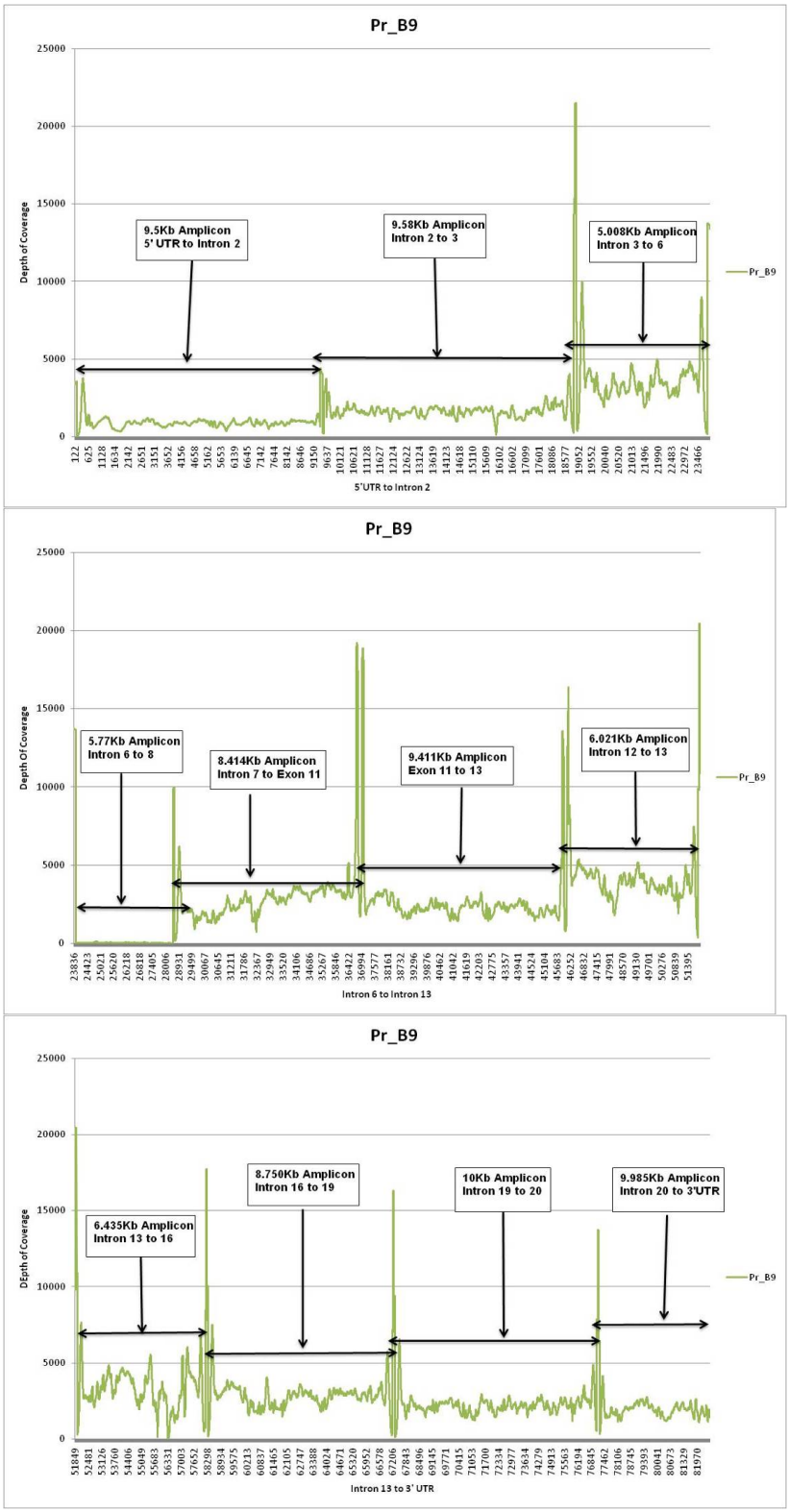


Figure 2.34 Coverage data for one patient sample. *These diagrams show the coverage data for one patient sample*

One amplicon appears to show very low coverage across most samples. This is the region exon 6 to exon 8. Figure 2.34 illustrates this region in the middle graph labelled 'intron 6 to intron 8'.

2.15.10 Coverage statistics

Table 2.15 coverage statistics

Sample	Coverage >30x
001	94.5%
002	96.5%
003	95.4%
004	94.3%
005	95.2%
006	95.2%
007	95.2%
008	94.9%
009	95.3%
010	67.3%
011	86.0%

Table 2.15 coverage statistics. *This table shows the coverage statistics for each of the 11 samples sequenced.*

Coverage statistics show the percentage of the target region covered for each sample (Table 2.15). 9/11 samples show coverage of more than 30X for over 94% of the target region. For two samples coverage of more than 30X was lower at 86% and 67.3%.

2.15.11 Phred scores

Phred scores are assigned to each base call giving a quality score for each base. Scores are all over 50; this demonstrates a probability of a base call being incorrect as 1:100,000.

2.16 Discussion

2.16.1 Detection of blinded causative mutations in 12 patient samples

This study uses LR-PCR and the Illumina GAII to sequence the whole of *BRCA1* gene in 12 patient DNA samples demonstrating that, in principle, NGS can be used in the clinic, to detect deleterious mutations in cancer susceptibility genes. The entire genomic region of *BRCA1* is sequenced in 11 samples (1 failed library preparation) on Illumina GAII platform using a multiplex sequencing protocol that pooled and sequenced all samples on one lane of a flow cell. LR-PCR is found to be an appropriate target enrichment method for the size of target region and the number of samples.

9 small insertions and deletions (indels) and 1 nonsense mutation are accurately detected in the correct samples. The 9 indels result in a frameshift mutation. Frameshift mutations result in an alteration of the gene's reading frame and a shift in the grouping of bases that leads to an altered or non-functioning protein product. The size of indels ranges from 1bp to 4bp. The nonsense mutation results from a single base substitution in which there is an amino acid change from Glutamic Acid to a Stop Codon. The reason the one 11bp indel is not detected in one of the samples may be because the alignment software is not able to detect indels of this size. Morgan et al (2010) note that the alignment and variation software program CASAVA does not detect insertions or deletions. Capillary sequencing is employed to investigate this sample for an indel in exon 2. This demonstrates that the sample is positive for the 188del11 mutation in exon 2. This may be remedied with the use of alternative software (NextGENe) or with the use of paired-end sequencing, which should improve accuracy in base calling. Numerous single base changes are detected; many of these are missense mutations that result in a substitution of one amino acid for another, however the functional impact of these is not assessed. In the next study this limitation is addressed with the use of functional prediction software programs. 15 coding non-synonymous single base changes are identified; 10 coding synonymous single base changes are detected. A very large number of unclassified variants are detected in the regulatory and intronic regions of *BRCA1*. These include indels of 1bp to 5bp and single base substitutions. Many of these are repeated in several samples and some are unique to specific samples, however, the functional effect of these cannot easily be assessed; indeed many may be sequencing artefacts. Moreover, as these samples are positive for deleterious coding mutations, it is likely that these are of little significance here. Nevertheless, detecting these intronic changes is important as this

shows that the methodology used here is able to identify small changes in a very large genomic region.

2.16.2 Scaling up

In order to increase the scale of targeted re-sequencing to that required in a diagnostic laboratory or for use in very large scale direct sequencing of candidate cancer susceptibility genes, a number of issues need to be addressed. There are a number of areas where bottlenecks in the protocol create an obstacle in increasing throughput. Overcoming these will ensure streamlined procedures and allow for scaling up. This study identifies the first bottleneck prior to library preparation during normalisation of the PCR products. For 12 patients for one gene it involved 132 different PCR products. If this is calculated for 48 samples for 3 genes (2 at 11 fragments and 1 at 7 fragments), it will require 1392 PCR products. The PCR amplification can be automated which could reduce the risk of pipetting errors or contamination. The purification of PCR products prior to quantitation for normalisation and pooling can be done in 96 well plates using SPRI beads. The protocol for this is outlined as 'supplementary protocol 1 in the paper by Mamanova et al (2010) in which an alternative library preparation protocols is described. The quantitation of PCR products for normalisation would require automation to reduce the risk of pipetting errors. Other methods of quantitation and normalisation will be investigated, for example, PicoGreen or qPCR.

It took several days to prepare 12 libraries; therefore, preparing 48 will take around 3 weeks. There are new automated systems available for library preparation. For example, SPRIworks Fragment Library System by Beckman Coulter Genomics advertises that 20 libraries can be prepared in just one day compared to 4 libraries in one day when done manually. Automated systems obviate the need for certain steps in the current protocols, such as size selecting by cutting out gels. As throughput requirements increase the cost of sequencing may reduce, however, the cost of the library preparation will inflate with the use of more reagents and indexes for multiplexing.

The first step in library preparation is fragmentation of DNA samples. This step, performed using the Bioruptor to sonicate samples will take 2-3 days to fragment 48 samples. Using a Covaris AFA (Adaptive Focused Acoustics) machine to fragment samples results in improved more uniform fragments and removes the need to size select so that 48 samples would take 2 hours (Mamanova et al 2010).

The purification steps required between each step in library preparation could be done using QiaQuick 96 well plates rather than the spin columns used in this pilot study; this will improve the procedure, save time and be more streamlined. 96 well plates will also be used for the purification steps in the next studies.

2.16.3 Coverage uniformity

Sensitivity and specificity in base calling is achieved with sequencing coverage uniformity. However, variation in sequence uniformity has been identified in previous literature when using LR-PCR as a target enrichment strategy (Harismendy & Frazer 2009). In this study, depth of coverage is highly variable both between and within samples. There are a number of reasons why coverage is uneven including normalisation, design of multiplexed library preparation protocols and the LR-PCR itself.

If normalisation of LR-PCR fragments is not calculated accurately then this could lead to biases in sequencing coverage. In this study, quantitation of DNA concentration in PCR products is measured individually using the Qubit Fluorometer, and as such may not be accurate enough. An alternative quantitation method that could be automated for scaling up in the next follow on project may be more appropriate, such as quantitative PCR (qPCR). It is difficult to assess the effect of the omission of one sample in the final pooling; it is possible that this may have a negative effect on coverage uniformity.

Mamanova et al (2010) note that tagging fragmented DNA with indexes for bar-coding results in uneven coverage and their protocol for preparing libraries in 96-well plates includes the index sequence in a mid-region of the reverse PCR primer at the PCR enrichment stage. Another modification to the protocols to achieve improved uniformity is to use 5' blocked primers. Harismendy & Fraser (2009) suggest that using LR-PCR as target enrichment produces an overrepresentation at the ends of amplicons, specifically a 50bp region at the 5'ends. They suggest that the use of 5' blocked primers improves coverage uniformity.

Coverage of 30X is considered to be the minimum for accurate base calling. Very high coverage could also cause difficulty in base calling as some software programs apply filters that effectively exclude bases in areas of very high coverage. One crucial point in depth of coverage is highly accurate quantitation of DNA concentration in PCR

products and thus accurate normalisation. In the results section coverage statistics are outlined; 9 of the 11 samples show >30X coverage of >94% of the amplified sequence.

One amplicon (intron 6 to intron 8) shows low coverage across the majority of samples. This could be due to incorrect calculation of DNA concentration for normalisation purposes or due to the region being rich in A-T repeat sequences. Harismendy et al (2009) make comparisons between the three main NGS platforms. They advise that the SOLiD and Illumina systems both show areas of low or no coverage and that these regions are noted for their A-T rich repeat sequences. The clear benefit of improving uniformity of coverage is that it increases depth of coverage throughout the entire target region. Therefore, refinement of library preparation protocols will not only increase throughput, but will also improve quality control.

2.16.4 Single-read vs. paired-end read sequencing data

Single read (SR) sequencing data is likely to show marginally less accuracy when compared to paired-end sequencing (PE) data. PE data is advanced in terms of precision and accuracy in detecting insertions and deletions. The read length generated via PE sequencing is clearly twice as long as that of SR. This will be crucial if sequencing larger genomes or those containing large regions of repeat sequences, since it will be very difficult to align data accurately through these if using short SR data. For example, if the short SR data sequence appears in several areas in the gene and if this sequence has a mutation within it, then it will not be possible to ascertain the position of the mutation. With PE sequencing the two read sequence pairs are a known distance apart, thus ascertaining their position should be more possible. In addition, if the single read sequence includes repeats, then this read may be aligned anywhere within the repeat sequence. Since PE reads essentially double the read length then it is more likely that, at least, a portion of the reads will include regions outside of these repeat sequences.

In order to detect larger indels it will be necessary to use PE reads. In this pilot study single reads of 76bp are performed. It is known that *BRCA1* comprises 41.5% *Alu* repeat sequences, with one *Alu* repeat sequence occurring approximately every 650bp throughout the entire sequence of the gene (Tancredi et al 2004). In order to detect larger insertions and deletions it will be imperative to use PE sequencing. If these *Alu* sequences are 69-231 bp in length (Smith et al 1996), and the SR length is just 76bp then it is clear that this could be problematic and that aligning some 76bp reads against the reference genome sequence would be almost impossible. With 2 x 100bp (the new HiSeq2000) paired reads that are a known distance apart (i.e. insert sizes of ~300bp)

this will make alignment to the reference sequence much more precise. It may be possible to improve accuracy in alignment of reads by reducing the selected fragment size, however, this could increase the time it will take to analyse these data.

Using overlapping PE reads could further increase the accuracy of sequencing data produced. This is achieved by size selecting fragments of a smaller size than the combined read length. This is probably best accomplished using the longest read lengths available with the Illumina GAII. The additional read depth attained with PE sequencing will also increase coverage across the region sequenced, thus increasing the accuracy of mutation detection. Finally, the new HiSeq2000 system will also increase coverage depth. This system has a dual flow cell in which reads are taken from both sides of the flow cell, effectively doubling the number of reads per flow cell. This system is capable of the slightly longer read length of 100bp, which will increase accuracy of alignment further.

2.16.4 Quality controls

Quality controls are in place at all stages. These include the quantitation and validation of libraries using the Agilent Bioanalyzer 2100; the cluster density report produced by the Genome Analyzer; coverage data produced by analysis software; and Phred scores for base call quality produced by the analysis software. Phred scores are assigned for each base call and this system demonstrates that quality is in the region of 50-60 for each base call; which signifies that there is a probability of 1:100,000 that a base call is incorrect.

2.16.5 Cost comparison of sequencing methods

Table 2.16 shows a comparison of sequencing costs for different sequencing methods (2010 costs); these are for the current method used in the genetics clinic (Sanger sequencing) for diagnosis of *BRCA1* or *BRCA2* mutations and the method used in this study (LR-PCR and NGS). This table also gives an estimated cost if the next study were to sequence 48 samples in 3 genes. The Illumina costs are based on one lane of a flow cell and the current research costs are based on PCR and capillary sequencing at the current (2010) price. The current clinical costs are derived from the cost charged by NHS centres and this price includes MLPA for both genes.

Table 2.16. A comparison of costs of sequencing methods

	Current Clinic	Current Research	Pilot Study 12 samples. BRCA1 only GAII	Follow-on 48 samples. 3 Genes. HiSeq2000
	BRCA1/2	*BRCA1/2*		
PCR		£1 x 44 + 5%	£1 x 12 + 5%	£1 x 29 +5%
		£46.20	£138.60	£1,461.00
Lib Prep			£250 x 12	£250 x 48
			£3,000.00	£12,000.00
Sequencing				
	£1,000.00	£217.80	£1,330.00	£4,000.00
Total cost per sample	£1,000.00	£264.00	£372.38	£363.77
Total cost per gene	£500.00	£132.00	£372.38	£121.25
Time to sequence only				
No. Samples	N/K		12	48
Region covered	coding only	Coding only	regulatory regions, coding, inronic	regulatory regions, coding, inronic

Table 2.16 A comparison of sequencing methods costs. *This table compares the sequencing costs for different sequencing methods (2010 costs).*

2.16.6 Rejecting the use of Long Range PCR as target enrichment method

LR-PCR is a highly effective method of enriching a large genomic region in a small number of samples. However, it is debateable that the inclusion of all regions of genes is necessary in order to assess risk. Deleterious changes in *BRCA1* or *BRCA2* genes are generally considered to be those that result in a non-functioning protein product. Therefore, those that reside in non-coding regions or that do not result in amino acid changes in coding regions cannot be classified in terms of their influence on breast or ovarian cancer susceptibility. In addition, in the clinical setting both variants of uncertain significance and non-coding changes are currently ignored.

The discovery of genetic alleles for ovarian cancer will require interrogating very large sample sets; therefore, increasing throughput is of paramount importance. Thus, it is considered prudent to change the experimental design to only include a specific highly targeted region (the coding regions) of the genome that are most likely to be relevant in protein function. These coding regions can be directly related to the clinical setting as their probable pathogenic impact can be more accurately assessed.

2.16.7 Dealing with technical sequencing artefacts

In the studies that follow on from this pilot study, issues in distinguishing technical artefacts from real genetic changes need to be addressed to improve specificity and sensitivity in mutation detection. This will be addressed in two ways: firstly, in the use of paired-end sequencing which results in improved alignment and secondly, in the

data analysis steps in filtering out multiple reads and likely variants using computational methods and improved data analysis pipelines.

2.17 Conclusion

To conclude, all known mutations are identified using LR-PCR and sequencing on the Illumina GAII platform with the exception of one 11bp indel and one causative mutation in the sample that failed at library preparation stage that could not be sequenced. The accuracy and sensitivity of mutation detection using this protocol is dependent upon experimental design of sequencing as well as improvements in bioinformatics and data analysis. The use of paired-end sequencing to improve read mapping is a fundamental step. Filtering out duplicate reads and devising methodology to discern and remove technical sequencing artefacts are necessary in order to conclude that Long Range PCR and multiplexed NGS for mutation detection are as accurate as PCR enrichment and Sanger sequencing. However, these issues are not insurmountable. NGS is a rapidly advancing technology with increasingly more streamlined sample preparation whilst simultaneously becoming cheaper and faster with increasing sequencing capacity. The cost and throughput capabilities of NGS mean that it could be used to test a wider population than are currently tested under NICE recommendations. The potential benefits of next generation sequencing approaches within research and diagnostics are vast. This thesis will develop next generation sequencing approaches still further. At the same time focus will be given to scaling up research to use this established technology in mutation detection and examine the frequency of both known and novel cancer susceptibility genes in large sample sets of cancer cases and controls.

Chapter Three

A high throughput targeted sequencing approach to evaluate the penetrance and prevalence of germline mutations in 6 DNA repair genes in epithelial ovarian cancer

3.1 Introduction

There is a strong rationale that gene variants, additional to *BRCA1* and *BRCA2*, of moderate to high penetrance exist for epithelial ovarian cancer, which can be discovered using massively parallel next generation sequencing approaches to scan the whole genome, or whole exome or targeted regions of the genome. It is known that the most significant factor for the development of epithelial ovarian cancer is family history of the disease. A single first-degree relative (FDR) affected by ovarian cancer confers an elevated disease risk in women two-three times greater than the general population (Stratton et al 1998). Lichtenstein et al (2000) suggest, from studies of twins, that the inherited component of ovarian cancer risk could be as high as 22%.

The highly penetrant genes, *BRCA1* and *BRCA2* are most often found in families where there are several cases of ovarian cancer and/or breast cancer. *BRCA1* mutation carriers show penetrance levels of 40% to 50% for ovarian cancer. *BRCA2* mutation carriers show penetrance levels of 20% to 30% for ovarian cancer (Ford et al 1998; Ramus et al 2007). Other highly penetrant genes identified include the genes associated with Lynch syndrome (the DNA mismatch repair genes), which show average cumulative lifetime (to age 70) penetrance levels of 6% to 12% for ovarian cancer (Aarnio et al 1999; Lu & Daniels 2013). Research conducted in the last few years identifies additional high-moderate penetrant genes in ovarian cancer (as well as breast-ovarian cancer familial syndromes). These include genes that function in the *BRCA1* network of DNA double strand break repair. Meindl et al (2010) detect rare variants in *RAD51C* and Loveday et al (2011) find deleterious variants in *RAD51D,* in ovarian cancer cases negative for *BRCA1* or *BRCA2*. A study by Rafnar et al (2011) finds *BRIP1* deleterious variants also result in an increased risk of ovarian cancer.

Whilst linkage studies have not revealed genes with the level of penetrance as high as *BRCA1* or *BRCA2*, recent sequencing studies are finding genes that result in a considerable increase in ovarian cancer risk (>10x more than those without variants in these genes). Estimates suggest that these highly penetrant genes comprise up to

40% of the familial ovarian cancer risk (Holschneider & Berek 2000). It is likely that the remaining, currently unidentified, susceptibility alleles for ovarian cancer are attributable to genes at the moderate to low penetrance level. This study will establish a novel system for library preparation, which will allow for scaling up to very high-throughput levels; and at the same time will identify both known and unknown epithelial ovarian cancer susceptibility genes. This will include validating recently discovered breast-ovarian cancer susceptibility genes and detecting novel candidate genes in a set of 5 interacting DNA repair genes, plus 1 stand-alone DNA repair gene.

Since around half of the inherited susceptibility to epithelial ovarian cancer is attributable to genes other than *BRCA1* or *BRCA2,* discovering these remaining rare variants is required to enable improved risk prediction and early detection of disease. Identifying these rare variants requires large scale sequencing studies of either a few targeted candidate genes or whole exome sequencing. This study uses a candidate gene approach to investigate the contribution of six DNA repair genes to epithelial ovarian cancer.

3.1.1 Technological advances in next generation sequencing approaches for mutation detection

The previous chapter established the requirement to address issues of library preparation bottlenecks in scaling-up research to fully utilise the capabilities of next generation sequencing technologies in mutation detection for identifying cancer susceptibility genes. This chapter introduces and evaluates a novel approach to library preparation demonstrating its use in large scale mutation detection studies in ovarian cancer.

3.1.2 Introducing a novel library preparation system – Fluidigm Access Array

Large-scale population based screening studies are required to discover novel ovarian cancer susceptibility genes. The main hurdle to increasing throughput is the bottleneck that exists at the library preparation stage. The cost of library preparation and the time to isolate the region of interest and prepare DNA libraries is the predominant limitation preventing the full use of Illumina sequencing capacity. Library preparation is time consuming and expensive. The Fluidigm Access Array system is a microfluidic platform that can perform target enrichment and library preparation simultaneously in a high throughput format. This system essentially removes this bottleneck reducing the time and cost in sample preparation for very high throughput re-sequencing projects. Expensive library preparation kits are not necessary with the Fluidigm platform and the

input DNA quantities are greatly reduced in comparison to genomic capture methods. At the same time, multiplexing at 384 samples uses the increased capacity of the HiSeq2000. This study uses the Fluidigm Access Array, which is a complete sample preparation system to combine target enrichment and library preparation into one platform.

Figure 3.1 Target enrichment using the Fluidigm Access Array system

Figure 3.1 Target enrichment using the Fluidigm Access Array system. *Fluidigm Access Array System is capable of preparing 48 samples with 48 PCR reactions simultaneously resulting in 2,304 reactions and final PCR product pools of 48 for each of 48 samples.*

The Fluidigm Access Array system is a method for target enrichment and library preparation that simultaneously prepares multiple targeted re-sequencing libraries using multiplex PCR on a microfluidic platform (Figure 3.1). The Access Array can be used to prepare 48 libraries in around 5 hours. 48 PCR amplicons are generated for each of 48 samples on the 48.48 Integrated Fluidic Circuit (IFC). At the same time each sample can be indexed to produce PCR products that are immediately ready for re-sequencing. This circumvents the time consuming and costly library preparation step and effectively eliminates the bottleneck at this stage noted in re-sequencing studies.

Figure 3.2 Overview of the Fluidigm Access Array protocol

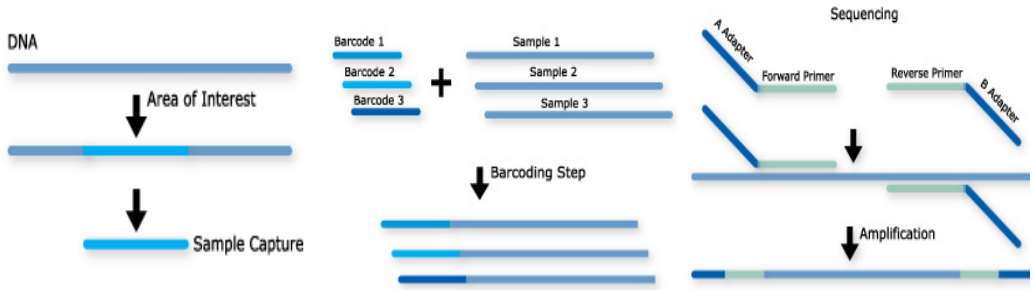


Figure 3.2 An overview of the Fluidigm Access Array protocol*. The required regions of interest (ROI) are identified and PCR primers are designed to include adaptor sequences and sample specific indexes to produce tagged amplicons ready for cluster generation on the Illumina platform. This essentially eliminates the library preparation step*

The design by Fluidigm included a wet validation of primers and a gel like image to demonstrate that the primers successfully amplify the regions of interest (Figure 3.3).

Figure 3.3 Wet-test amplification results

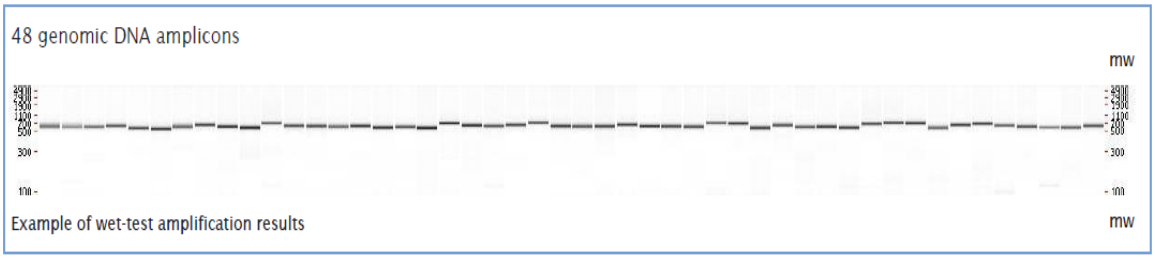


Figure 3.3 Wet-test amplification results*. This gel-like image shows the PCR reactions for each of 48 amplicons and demonstrates successful PCR reactions*

Forward and reverse primers are supplied in a 96-well plate format to enable ease and speed of multiplex PCR set up. Through the PCR reaction sample specific index primers and Illumina sequencing adapters are adjoined to PCR amplicons. Four primer sequences are used to prepare samples for sequencing with an Illumina HiSeq2000. The custom adapter sequence tag (CS1) is joined to the target specific primer sequence (TS) to form one primer sequence. The Illumina PE sequence tag (PE1) is also joined to the TS. The other two are the same, except reverse sequences. A fifth primer is needed (index SP) to read the index sequences. As sample specific indexes are adjoined the resultant 2,304 PCR amplicons are ready for pooling and multiplex sequencing on one lane of the Illumina HiSeq2000.

Figure 3.4: Overview of the 4-primer PCR protocol

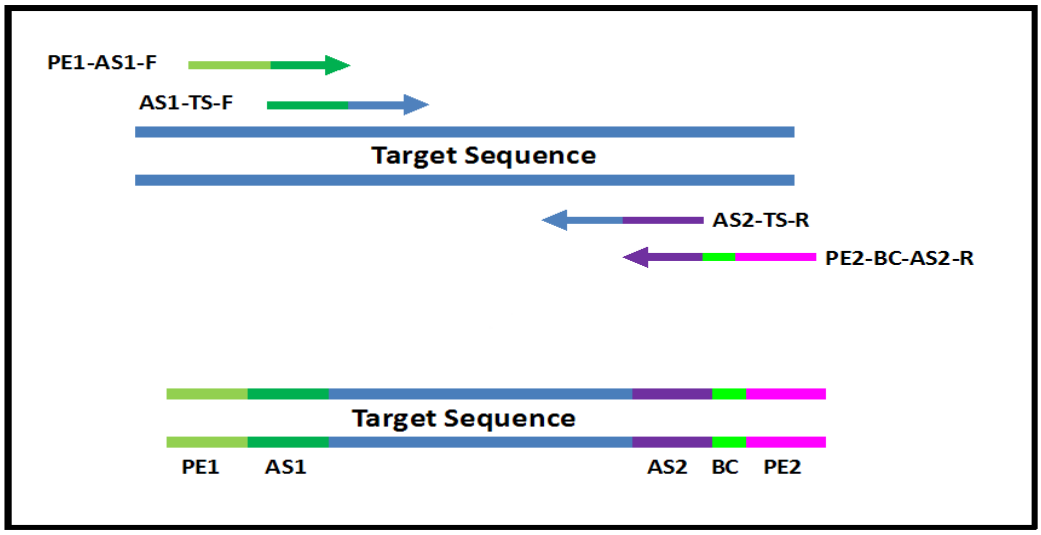


Figure 3.4 Overview of the 4-primer PCR protocol. *AS1 = Adaptor sequence tag 1; AS2 = Adaptor sequence tag 2; TS = Target-specific primer sequence; PE1 = Paired end sequence 1; PE2 = Paired end sequence 2; BC = Barcode sequence. Note that in this study custom sequencing tags (CS1 and CS2 were used because custom indices for 384 per lane multiplexing were used instead of standard 96 multiplexing indices).*

It is possible to multiplex further with the Fluidigm Access Array to include up to 10 primers in each well (10Plex). To do this successfully requires following specific rules on primer pair mixing. It is important to mix primer pairs that do not produce an overlapping PCR amplicon, such as primer pairs that overlap to amplify one larger exon. The mixed primer pairs should amplify regions that are at least 5 Kb apart. Primer sets should ideally amplify products that are all within 20% of the average pool size. For example, if the pool size average is 200 bp, then all primers should produce PCR products of 160 bp to 240 bp. Mixed primer pairs should produce amplicons of a similar GC content. Primers with multiple annealing sites should be avoided and primer pools are best tested for primer-dimer formation before use and avoided.

Multiplexing PCR reactions in each well results in an increase in sample throughput on the 48.48 Access Array. In a single plex reaction design the 48.48 IFC can prepare 48 samples and one PCR reaction for each sample, which results in 48 PCR reactions for each sample and a total of 2,304 PCR reactions per Access Array chip. Multiplexing at 3plex, on the 48.48, means that 3 PCR reactions are performed in each well for 48 samples and 48 wells. Thus, 144 PCR reactions are performed for each of 48 samples and a total of 6,912 PCR reactions can be achieved for each Access Array chip; resulting in an impressive reduction in time, labour and cost. Multiplexing at 3plex means that all amplicons to cover the whole coding region of the 6 candidate genes can be prepared for 384 samples in each chip which reduces the total number of chips

and reagents used as well as the time taken to prepare the chips by 3. The only possible caveat here would be whether these reactions could successfully be mixed in each well. If inappropriate amplicons are mixed then those that do not sequence will have to be repeated. This could make multiplexing prohibitive, as it would be necessary to re-amplify and then to re-sequence the failed products.

3.2 Recently discovered ovarian cancer susceptibility genes

The study by Meindl (2010) detects 6 heterozygous pathogenic mutations in *RAD51C* and reveals that these occur in 1.3% of affected cases from breast and ovarian cancer families compared to no pathogenic mutations in the unaffected controls or in those families with breast cancer only. A study, published during this project, by Loveday et al (2011) which analyses *RAD51D* for germline mutations in breast and ovarian cancer families finds 8 pathogenic mutations in cases and only 1 in the unaffected controls this study suggests that risk estimates for *RAD51D* mutations are 6.30 for ovarian cancer and 1.32 for breast cancer. In this way, looking at one gene at a time, discovering the remaining heritability to ovarian cancer will take an extremely long time; therefore, it is important to use technology that will analyse multiple genes in large sample sizes.

3.3 The research questions

1. Does the Fluidigm Access Array system offer a viable solution to addressing the bottleneck in library preparation?
2. Is Fluidigm microfluidic technology both an accurate and rapid method for sequencing library preparation in a large sample set for cases and controls? This will be assessed by evaluating time to prepare libraries, depth of coverage, evenness of coverage and mutation detection sensitivity.
3. Is this system able to characterise the mutation frequency of 6 DNA repair genes in ovarian cancer?

3.3.1 The research questions in context: how this research impacts on the health of the population.

Strategies that focus on early detection and improved risk prediction in ovarian cancer are being suggested as an effective clinical intervention in breast and ovarian cancer. Indeed, genetic testing is already in place in order to identify women at high risk of developing breast or ovarian cancer and these women are offered early screening or risk reduction surgery. Developing a technology that is capable of identifying additional

genes in epithelial ovarian cancer will have a direct translational impact by validating the means by which cancer susceptibility alleles can be found. At the same time, this novel approach that increases throughput by such an extent will undoubtedly result in a wider population of women being able to benefit from genetic screening for ovarian cancer.

Identifying risk genes and risk factors provides further insight into the causes of ovarian cancer susceptibility. This knowledge will be invaluable in the refinement of risk prediction strategies allowing for earlier detection and prevention of disease. These studies could indicate elusive biomarkers for ovarian cancer detection. The establishment of high-throughput NGS approaches and improvements in technology during the progress of this research will lead to a wider population of women being offered genetic testing and assessment of cancer risk. Discovering the cause for the remaining ovarian cancer susceptibility will result in more women being offered prophylactic surgery that reduces the risk of development of ovarian cancer by more than 90%. This will result in individual personalised care and treatment tailored to an individual woman's level of risk and allow for a greater informed choice (prophylactic surgery or early surveillance and monitoring) than is currently available.

3.4 DNA repair and cancer susceptibility

The DNA damage response is a molecular signalling kinase cascade that is initiated following detection of DNA damage. This response involves the transcriptional regulation of genes implicated in repair and replication of DNA. DNA damage includes double strand breaks (DSB), nicks, gaps and genetic changes that prevent the replication of DNA. *ATM* and ataxia telangiectasia and Rad3 related (*ATR*) code for protein kinases that detect DNA damage. ATM exists in an inactive dimer state and has a role of specifically detecting DSBs, where ATM separates into active monomers once recruited to subsequently phosphorylate downstream targets.

Double strand break DNA repair mechanisms include homologous recombination (HR), non-homologous end joining (NHEJ) and microhomology-mediated end joining (MHEJ). In addition to DSB, homologous recombination is also the mechanism for the repair of stalled replication forks and DNA interstrand cross-links. Mutations that result in an inactivation of genes involved in HR have been identified as increasing the susceptibility to certain cancers. The recent discoveries of novel genes (*RAD51C* and *RAD51D*) further support the view that additional rare variants with moderate penetrance exist. By re-sequencing candidate genes in large cohorts of ovarian cancer

cases and equal numbers of healthy matched controls it should be possible to identify the likely pathogenic variants in novel cancer susceptibility genes.

3.5 Study candidate genes – 6 DNA repair genes

Six candidate genes are selected for targeted re-sequencing in a large series of ovarian cancer cases and controls. The five interacting genes being analysed are RAD51 paralog B (*RAD51B*), RAD51 paralog C (*RAD51C*), RAD51 paralog D (*RAD51D*), X-ray repair complementing defective repair in Chinese hamster cells 2 (*XRCC2*) and X-ray repair complementing defective repair in Chinese hamster cells 3 (*XRCC3*). The five interacting genes (*RAD51B*, *RAC51C*, *RAD51D*, *XRCC2* and *XRCC3*) are involved in homologous recombination and the sixth one SLX4 structure-specific endonuclease subunit homolog (*SLX4*) is linked to homologous recombination and Fanconi Anaemia. These candidate genes are associated with *BRCA1* or *BRCA2* and are involved in the *BRCA1* network in double strand break repair. These genes may be associated with high-grade serous ovarian carcinoma as they are involved in homologous recombination (TCGA data 2011). DNA samples include both patients of unknown mutation status as well as those already identified as negative or positive for mutations in *BRCA1* or *BRCA2*. The rationale for choosing these candidates is explained below in Figure 3.5.

Figure 3.5 The interaction between the RAD51 associated proteins

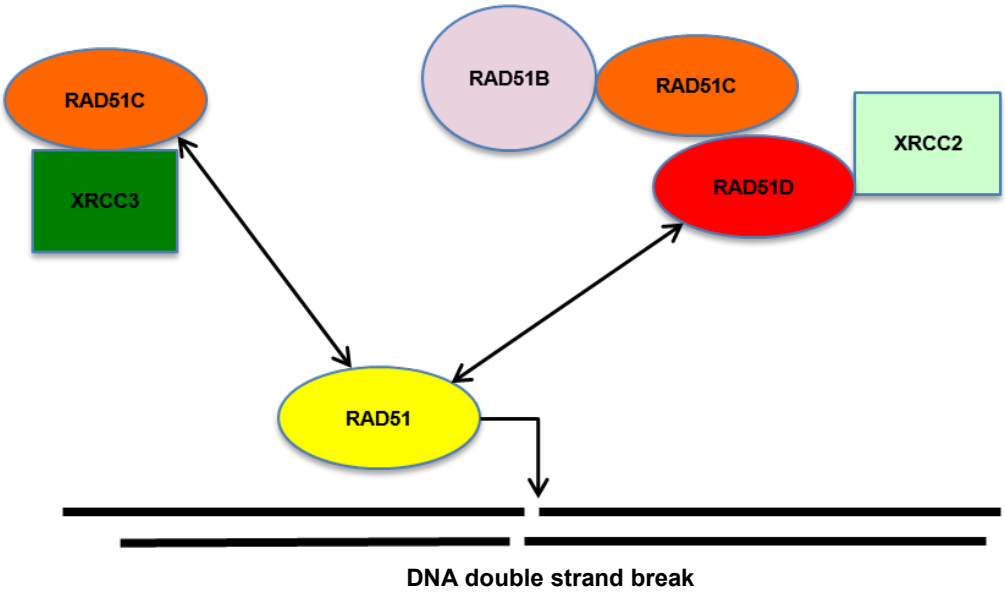


***Figure 3.5 The interaction between the RAD51 associated proteins**. Two distinct complexes are formed within this group of proteins. 1) RAD51C complexes with XRCC3; and 2) RAD51B, RAD51C, RAD51D and XRCC2 complex to form a complex known as BCDX2 complex. These complexes are recruited to DNA double strand breaks with BCDX2 functioning to bind gaps in one strand of a double strand DNA molecule.*

The RAD51 group of proteins (Figure 3.5) consists of RAD51 plus 5 RA51-like proteins RAD51B, RAD51C, RAD51D, XRCC2 and XRCC3. The primary function of this group of proteins is in homologous recombination repair of DNA double strand breaks. They form two complexes 1) RAD51C complexes with XRCC3 to form a heterodimer and 2) RAD51B, C, D, XRCC2 form a heterotetramer known as BCDX2; along with RAD51 both of these complexes play a role in strand invasion during homologous recombination (Miller et al 2004, Liu et al 2002).

Figure 3.6 Schematic representation of *RAD51B* (RAD51 paralog B) gene

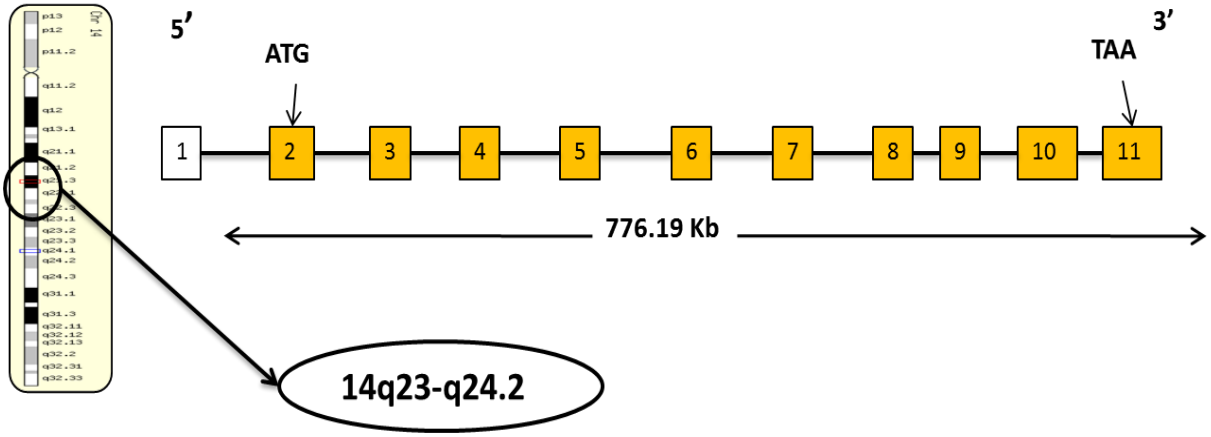


Figure 3.6 Schematic representation of RAD51B. ***RAD51B*** *is located at 14q24.1 Genomic co-ordinates: 14:68,286,495-69,062,737 (build GRCh37/hg19)*

RAD51B is involved in the hydrolysis of adenosine triphosphate (ATP) to adenosine diphosphate (ADP) as its structure (that includes nucleotide-binding motifs) implies that it is an ATPase. Expression of the protein encoded by *RAD51B* is observed in peak levels in the ovary, testis, thymus and spleen. Rice et al (1997) demonstrate that cells damaged by ionizing radiation express RAD51B, whereas cells that are not show no distinguishable levels of the protein.

Alternative splicing leads to the production of three transcripts, the largest of which is 2596 bp with an encoded protein product of 384 amino acid residues. RAD51B complexes with RAD51C; this heterodimer is involved in DNA double strand break repair via homologous recombination along with RAD51 and BRCA2. The protein product of *RAD51B* forms a complex with protein products of *RAD51C*, *RAD51D* and *XRCC2* to form a heterotetramer; this is known as the BCDX2 complex. BCDX2 binds gaps in one strand of a dsDNA molecule (Figure 3.5).

Figure 3.7 Schematic representation of *RAD51C* (RAD51 paralog C) gene

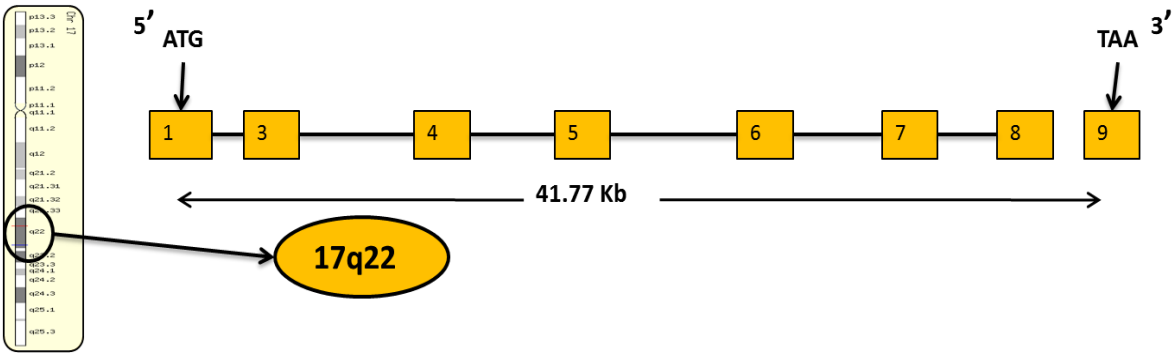


Figure 3.7 Schematic representation of RAD51C. *RAD52C is located at 17q22 genomic coordinates: 17:56,769,962-56,811,691 (build GRCh37/hg19)*

RAD51C protein forms a heterodimer with XRCC3, complexes with RAD51B and is critical in homologous recombination. Masson et al (2001) find that RAD51C and XRCC3 form a complex when co-purified and that the heterodimer only anneals single stranded DNA molecules. RAD51C is also part of the BCDX2 complex and is most likely to be involved in resolution of Holliday junctions. Liu et al (2004) find diminished Holliday junction resolvase activity in cells harbouring mutations in either *RAD51C* or *XRCC3*. The *RAD51C* transcript is 1322 bp producing a protein product of 376 amino acid residues with two isoforms. RAD51C assists in the phosphorylation of checkpoint kinase 2 (CHK2) which itself phosphorylates BRCA1 which in turn phosphorylates its downstream targets, such as p53 and Rb, which are involved in cell cycle control.

Figure 3.8 Schematic representation of *RAD51D* (RAD51 paralog D) gene

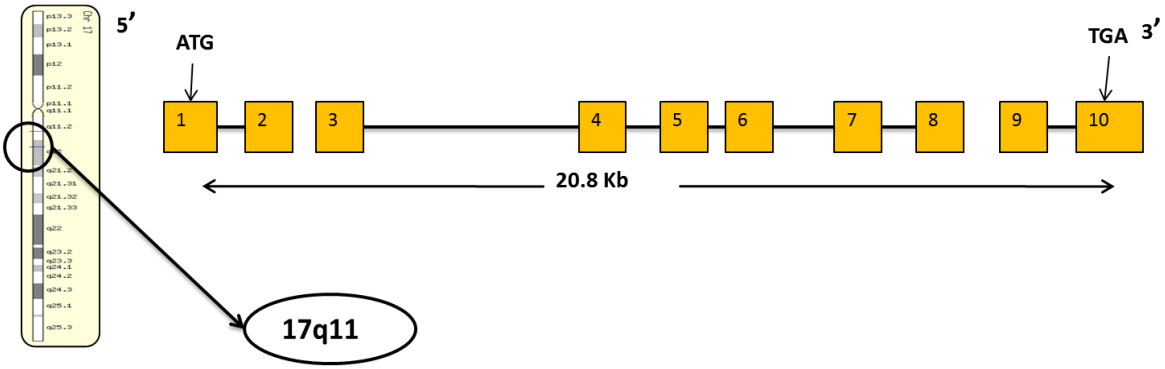


Figure 3.8 Schematic representation of RAD51D. *RAD51D is located at 17q12 genomic coordinates: 17:33,426,810-33,446,887 (build GRCh37/hg19)*

The protein encoded by *RAD51D* gene forms a complex with XRCC2 and is included in the BCDX2 heterotetramer, which binds gaps in one strand of a dsDNA molecule and as such has a role in homologous recombination. Several transcripts arise from

alternative splicing. The largest transcript is 2404 bp producing protein of 328 amino acid residues.

Braybrooke et al (2000) demonstrate that recombinant RAD51D hydrolyses ATP when Mg^{2+} is available. Braybrooke et al (2000) also note an association between XRCC2 and RAD51D. Loveday et al (2011), in a case-control study, detect 8 deleterious mutations in cases and 1 in the healthy age-matched controls. Similar to *RAD51C*, the incidence of these mutations is found to be higher in ovarian cancer than breast cancer; calculating the relative risk of ovarian cancer to be 6.3 (95% CI 2.86-13.85, $p=4.8 \times 10^{-6}$). Moreover, they demonstrate a higher prevalence of mutations amongst those cases with more than one case of ovarian cancer in their family. Interestingly, Loveday et al (2011) use functional assays to examine the effect of the loss of RAD51D function in tumour cells and their sensitivity to PARP inhibitors. Using short interfering RNA (siRNA) to silence *RAD51D* in tumour cells, they detect sensitivity to PARP inhibitors to be consistent with that noted in cells with non-functioning *BRCA2*. This suggests that PARP inhibitors may be valuable in the clinic for *RAD51D* positive patients along with those with *BRCA1* or *BRCA2* mutations.

Figure 3.9 Schematic representation of *XRCC2* (X-ray repair complementing defective repair in Chinese hamster cells 2) gene

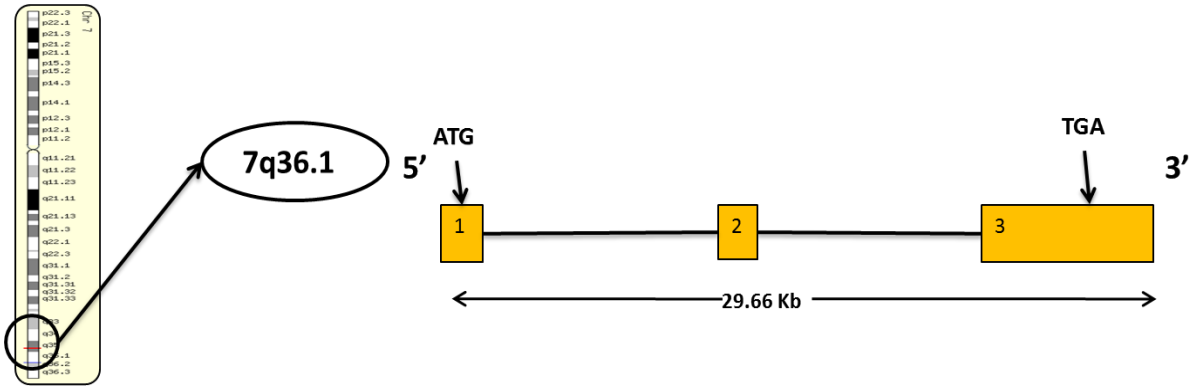


Figure 3.9 Schematic representation of XRCC2*. XRCC2 is located at 7q36.1 with genomic coordinates: 7: 512,343,582-152,373,249 (build GRCh37/hg19)*

XRCC2 has a role in the repair of DNA double strand breaks via homologous recombination (Johnson et al 1999). This was first noted in hamster cells that were lacking the protein encoded by the gene. The protein also has a role in the repair of chromosomal fragmentation, translocation and deletions. It is part of the BCDX2 complex, which binds gaps in one strand of a dsDNA molecule. The transcript is 3067 bp and the encoded protein is 280 amino acid residues.

Figure 3.10 Schematic representation of *XRCC3* (X-ray repair complementing defective repair in Chinese hamster cells 3) gene

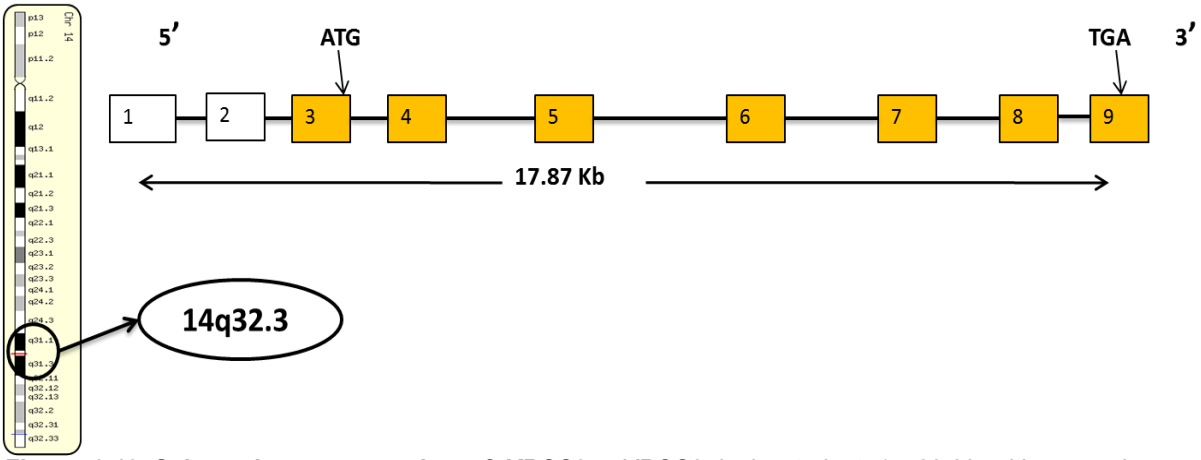


Figure 3.10 Schematic representation of XRCC3. *XRCC3 is located at 14q32.33 with genomic coordinates: 14:104,163,953-104,181,822 (build GRCh37/hg19)*

XRCC3 has a role in the repair of DNA double strand breaks via homologous recombination. In addition, the encoded protein has a role in the repair of chromosomal fragmentation, translocation and deletion. The protein product that is encoded by this gene interacts with RAD51 in homologous recombination (Liu 1998). XRCC3 also forms a complex with other proteins including BRCA2, FANCD2 and FANCG (Wilson 2008). There are two transcripts, 2622 bp and 2439 bp that arise from alternative splicing producing a protein of 346 amino acid residues.

Figure 3.11 Schematic representation of *SLX4* (SLX4 structure specific endonuclease subunit) gene

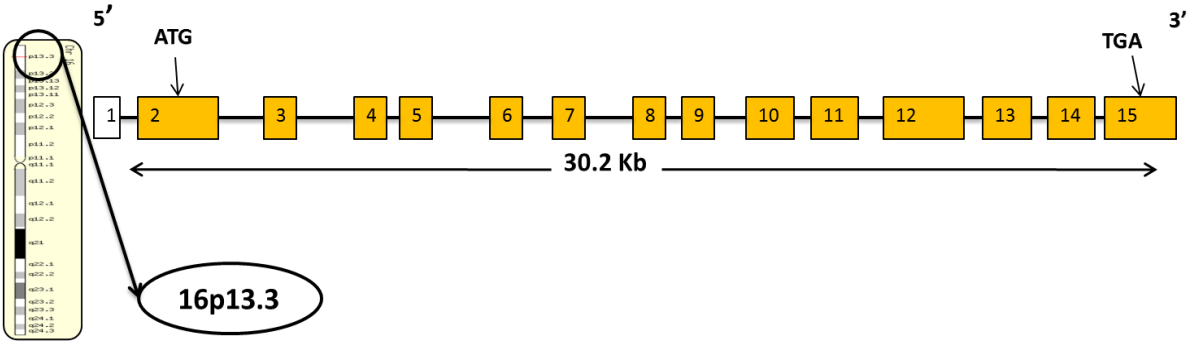


Figure 3.11 Schematic representation of SLX4*. SLX4 is located at 16p13.3 with genomic coordinates (GRCh37): 16:3,631,181–3,661,584(build GRCh37/hg19)*

SLX4 (SLX4 structure-specific endonuclease subunit) is previously known as BTBD12 (BTB/POZ domain-containing protein 12). Svendsen et al (2009) report the multiprotein complex that includes SLX4 is necessary for the repair of particular

categories of DNA damage and is key in resolving replication fork failure. SLX4 is a scaffold protein key in the formation of a multiprotein complex of enzymes required in the repair of DNA.

Cleavage of DNA strands and subsequent re-joining during DNA repair processes is necessary; this is coordinated through a number of structure specific endonucleases. SLX4 is an identified downstream target of ataxia telangiectasia mutated and ataxia telangiectasia and Rad3 related (ATM/ATR) (Svendsen et al 2009), both are protein kinases that have roles as DNA damage sensors and as such are pivotal in the initiation of the DNA damage response.

The protein product of *SLX4* gene is a regulatory subunit that co-ordinates the activity of various structure-specific endonucleases by increasing their activity. *SLX4* has a clear role in genome protection in DNA repair. SLX4 forms multiprotein complexes with excision repair cross-complementing rodent repair deficiency, complementation group 4-1 (*ERCC4-ERCC1*) and *MUS81* structure-specific endonuclease subunit (*MUS81*) and essential meiotic structure-specific endonuclease subunit (*EME1*) and *SLX1* (SLX1 structure-specific endonuclease subunit). Moreover, the protein interacts with telomeric repeat binding factor 2 (*TERF2*) and its associated interacting protein *TERF2IP*. Other protein interactions include protein polo-like kinase 1 (*PLK1*) and mismatch repair proteins (*MSH2-MSH3*). These other endonucleases have roles in repairing specific types of DNA damage that arise through homologous recombination and repair. For example, *SLX4* assists in the resolution of Holliday Junctions (HJ) produced during homologous recombination. *SLX4* is also involved in repair of stalled replication forks and acts as a docking platform for the construction of various structure specific endonucleases. *In vitro, SLX4* resolves Holliday Junctions (HJ) and *in vivo* they play a role in the repair of double strand breaks in homologous recombination (Svendsen 2009). *SLX4* has 15 exons, 14 of which are coding exons. Translation start site is in exon 2. The *SLX4* transcript is 7307 bp in length and its protein product is comprised of 1834 amino acid residues. During the progress of this project, researchers in Spain (de Garibay et al 2012) perform mutation screening in breast and ovarian cancer patients screen negative for *BRCA1* or *BRCA2*. They use high resolution melting analysis to scan the full coding region of *SLX4* along with the flanking sequences to search for germline variants in a cohort of 486 breast or ovarian cancer cases. They conclude that loss of function mutations in *SLX4* are very low in non-*BRCA1* or *BRCA2* breast and or ovarian cancer families.

The FA complex is recruited to sites of interstrand cross-link for ubiquitination by FANCD2 and FANCI; together these proteins localise to DNA damage sites along with BRCA2 (FANCD1), BRIP1 and PALB2. The following stage of repair is HR which involves RAD51 and associated proteins, as previously described.

Figure 3.12 A schematic representation of the FA-BRCA DNA repair pathway

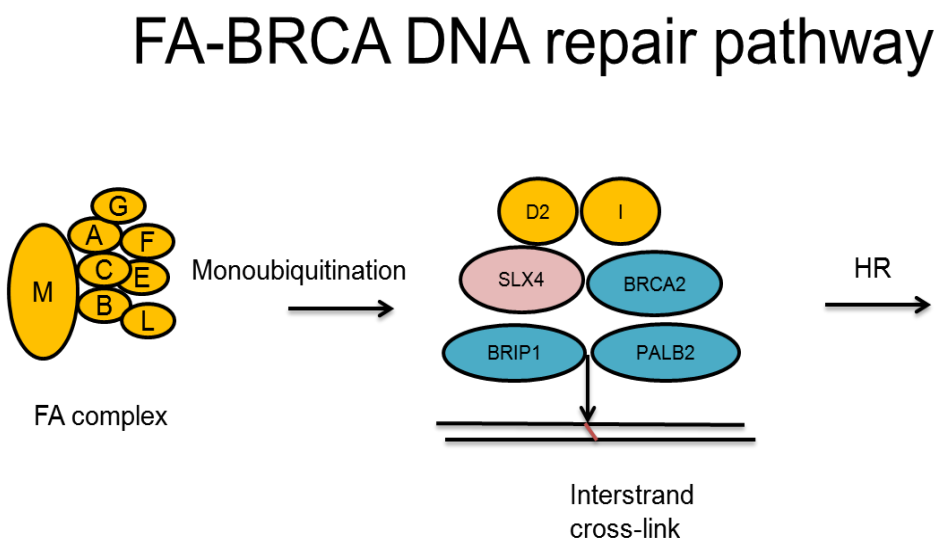


Figure 3.12 A schematic representation of the FA-BRCA DNA repair pathway *re-drawn and adapted from Levy-Lahad (2010) Fanconi anaemia and breast cancer susceptibility meet again. Nature Genetics. Vol 42. No. 5. 368-369*

3.6 Summary table of samples for next generation sequencing

3.6.1 Samples in the study

The cases and controls included here are drawn from a number of different sources internationally. They are from a population based case control series and two familial ovarian cancer registries and a polish familial cancer case study. The table overleaf (Table 1.1) gives details of the origins of samples.

Table 3.1 Summary of samples sourced from ovarian cancer studies or familial ovarian cancer registries

Registry	FH/Genetic/Histology	No. Cases	No. Controls
GRFOCR (Glider Radner Familial Ovarian Cancer Registry)	All (except 20) screened negative for *BRCA1/2*. All have FH (>2 OC) and Serous histology	175	
MALOVA (Malignant ovarian cancer study)	FH	44	192
	Serous histology	192	
UKOPS (UK ovarian cancer population study)	FH + Serous histology	579	384
UKFOCR	> 1 FDR with ovarian cancer	50	
POL (NCI ovarian case control study in Poland	> 1 FDR with ovarian cancer	96	
JAC Polish ovarian cancer study	Polish healthy control samples		96
AOCS (Australian Ovarian Cancer Study)	Serous histology	421	460
Totals		1557	1131

Table 3.1 Summary of samples sourced from other ovarian cancer studies or familial ovarian cancer registries. FH = family history, FDR = first degree relative

All of these samples are screened negative for mutations in *BRCA1* or *BRCA2*, except for 336 cases in UKOPS population study. Therefore, the purpose of re-sequencing these samples is to answer the question "do these candidate genes account for a proportion of familial susceptibility to ovarian cancer?" This study will aim to identify additional ovarian cancer susceptibility genes that could result in new routine tests being available in the clinical setting. If the risks are calculated to be high, then this could impact upon the management of ovarian cancer risk to identify rapidly and cost efficiently those women at increased risk of developing ovarian cancer. This is a case-control study, in which the controls are women unaffected by ovarian cancer, and age

matched to the cases. At the beginning of this study, only *RAD51C* gene is known to have mutations; this study includes 6 blinded *RAD51C* positive controls as a means of assessing the sequencing approach.

3.6.1 Gilder Radner Familial Ovarian Cancer Registry (GRFOCR)

This registry was established in 1981 in the US with the aim of researching novel genes linked to familial ovarian cancer. This self-referral registry includes families containing at least 2 ovarian cancer cases. The registry is based within the Roswell Park Cancer Institute Gynaecologic Oncology Department and collates familial cancer information.

There are four main aims of the GRFOCR:

1) To collect family history data from ovarian cancer patients within families with at least 2 cases of ovarian cancer or individuals with a related cancer syndrome.
2) To record the incidence of cancer, by referring to pathologists and medical records of tissue samples.
3) To build a bio bank of biological samples drawn from volunteers in the registry.
4) To allow the use of these biological samples for approved research.

Volunteers in the registry are selected if they meet at least one of the criteria listed here:

1) At least 2 cases of ovarian cancer in the family.
2) 1 ovarian cancer plus 2 other cancers
3) At least 1 woman in the family with at least 2 primary cancers, one of which must be ovarian cancer
4) At least two cases of cancer in the family with 1 being ovarian and the other diagnosed under 45 years (early onset).

Volunteers must also sign consent forms agreeing to allow access to medical records; family history data is obtained via written forms. Once volunteers are accepted as members into the registry they complete blood forms for giving blood samples. The registry requests permission to contact the relatives of registry members and these are invited to take part. More than 2,616 families are part of GRFOCR and 2,011 of these families contain 2 or more cases of epithelial ovarian cancer. The whole registry

includes 4,987 women whom have had epithelial ovarian cancer (Piver et al 1984, Greggi et al 1991)

3.6.2 UK Familial Ovarian Cancer Registry (UKFOCR)

The UK Familial Ovarian Cancer Registry (UKFOCR) is a UK based study originally known as the UK Co-ordinating Committee for Cancer Research (UKCCR) that began in 1991. The registry includes information on 391 ovarian cancer families containing at least 2 cases of epithelial ovarian cancer in first or second-degree relatives. Within these families there are 1,433 cases of cancer.

For both GRFOCR and UKFOCR, information gathered includes family history of cancer, reproductive records and other medical data, including medical history, use of oral contraceptives and/or hormone replacement therapy (HRT); blood as well as breast and ovarian tumour tissue samples are also stored for a proportion of the registry members.

3.6.3 Australian Ovarian Cancer Study (AOCS)

AOCS was established in 2001 and is a collaborative effort between institutions within Australia including University of Melbourne, Queensland Institute of Medical Research, Peter McCallum Cancer Institute Melbourne and Westmead Hospital Sydney. This population-based case-control study recruited women throughout Australia with a diagnosis of invasive or low malignant potential epithelial ovarian cancer between January 2002 and June 2005. All women are aged between 18 and 79 years at recruitment. The study consists of 2,319 cases of epithelial ovarian cancer and 1,509 healthy age-matched controls. Full epidemiological data are held on study participants to include general health and lifestyle, family history of cancer, medical history data and details on reproductive history (Merritt et al 2008)

3.6.4 UK Ovarian Cancer Population Study (UKOPS).

UKOPS is a UK case-control bio-banking study (Principal Investigators, Usha Menon, Simon Gayther and Ian Jacobs) with three mains aims:

1. To identify moderate penetrance genes for epithelial ovarian cancer using SNP analysis.
2. Evaluation of biomarkers for ovarian cancer cases, including pre- and post-surgery and subsequent to chemotherapy.

3. Classification of common medical conditions and symptoms in women with ovarian cancer compared to those without.

Criteria for inclusion as cases in UKOPS:

1. Women diagnosed with a primary invasive ovarian cancer or fallopian tube carcinoma.
2. Women diagnosed with a borderline ovarian cancer
3. Women pre-surgery diagnosed with an adnexal mass that may be ovarian cancer
4. Women not having surgery for a likely ovarian cancer

Criteria for inclusion as control in UKOPS:

1. Women that are part of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) that are undergoing annual screening and appearing healthy.
2. Women with a benign or borderline adnexal mass

The study (two years in duration) includes 10 UK Gynaecological Clinics. Follow-up is being conducted for 15 years and is due to end in 2020. Blood and tissue samples are stored with medical records of surgery, chemotherapy and histology reports. Health questionnaires are sent to participants periodically throughout the follow-up (Institute for Women's Health, UCL website accessed 02-08-2013)

3.6.5 Polish ovarian cancer studies

The Polish Ovarian Cancer Case Control Study (NCI POL) and the Polish Ovarian Cancer Study (JAC) are studies sponsored by the National Cancer Institute (NCI) in the USA. The National Cancer Institute (NCI) Polish Ovarian Cancer Case Control Study is a population based case-control study and includes ovarian and endometrial cancer cases diagnosed in women from Warsaw and Lodz in Poland between 2001 and 2003. Blood, urine and tumour samples are stored for each participant and the control women are healthy women age and location matched in proportion. The study includes 347 invasive ovarian cancer cases and 555 endometrial cancer cases. 2,798 healthy controls are included for breast, ovarian and endometrial cancer cases. DNA from tumours is also extracted and stored ready for appropriate studies (NCI Division of Cancer Epidemiology and Genetics website accessed 02-08-2013). The Polish ovarian cancer case control study is a hospital based case control study, which includes 423 cases and 417 healthy age-matched controls. Both of these studies examine the molecular epidemiology of Polish breast, ovarian and endometrial cancer and includes family history data, histology, stage and grade (Permuth-Wey et al 2013).

3.6.6 Malignant Ovarian Cancer Study (MALOVA)

The MALOVA study is conducted in Denmark and involves a multidisciplinary approach to the study of ovarian cancer including, molecular biology, epidemiology and biochemistry with the primary goal to ascertain accurate risk prediction and prognosis of the disease. Several Gynaecological centres throughout Denmark are included. Cases are women between 35 and 79 years with recruitment prior to surgery for suspected ovarian cancer identified between December 1994 and May 1999. The study includes 698 ovarian cancer cases, 219 ovarian borderline tumours and 450 benign ovarian tumours. Blood is taken prior to surgery and tumour samples stored from surgical procedures. All tumours have histopathology classification reviewed separately by two pathologists. International Federation of Gynaecology and Obstetrics (FIGO) staging of tumours, which is the standard staging for ovarian cancer, is recorded for each case (Hogdall et al 2003).

3.7 Study designs in population based genetic studies

This study is a population-based case-control study, which employs a candidate gene approach to identify gene variants for epithelial ovarian cancer. The genomic DNA of individuals, unrelated to each other, is analysed for changes in 6 DNA repair genes. The frequency of gene variants in cases is compared to the frequency observed in controls. This is in contrast to family-based designs in which families are analysed and controls within families are unaffected family members. Family-based studies are useful for the identification of highly penetrant genes within high-risk families; where population-based studies are useful for finding alleles with a low to moderate increase in risk in a wider population (Daly & Day 2001). Population-based design has the clear advantage that collecting samples is simpler and more likely to produce the larger samples sizes required to increase statistical power (Li et al 2010).

3.8 Research aims

The four main aims of this study are:

1. To establish a novel approach for target enrichment for high-throughput targeted NGS; this will represent an increase in scale compared to the pilot study, by using the increased capacity of Illumina HiSeq2000 technology.
2. To use this novel approach to characterise the mutation prevalence in 2,688 subjects from ovarian cancer case-control studies for 5 interacting genes in the DNA double strand break repair pathway (namely, *RAD51B*, *RAD51C*, *RAD51D*, *XRCC2* and *XRCC3*) that interact with *BRCA1* and *BRCA2*

3. To use NGS to characterise the mutation prevalence in the same 2,636 subjects of 1 additional DNA repair gene linked to Fanconi Anaemia in epithelial ovarian cancer (namely *SLX4*).
4. To evaluate the suitability of the Fluidigm Access Array platform and NGS in the clinical diagnostic genetics laboratory.

The fundamental aim of this study is to expand on the pilot study to re-sequence 6 DNA double strand break repair genes. A high-throughput population based case-control study design is being used and includes DNA samples from a variety of international ovarian cancer epidemiological case-control studies and familial ovarian cancer registers. Samples are identified that have previously tested negative for *BRCA1* or *BRCA2* mutations and some are sourced from patients with high-grade serous adenocarcinoma of the ovary (HGSOC).

3.9 Hypotheses under investigation

1. Fluidigm Access Array platform and highly multiplexed Illumina sequencing technologies are accurate, affordable and rapid methods for identifying genetic alleles in cancer susceptibility genes.

2. This novel method of mutation detection will be able to assess the mutation frequency in 6 DNA repair genes in a large series of ovarian cancer cases and controls.

3.10 Results

(Refer to Chapter 6 Materials and Methods page 255)

3.10.1 Study Design - A population based case-control study

This population based case-control study uses a novel highly multiplexed DNA sequencing approach to identify ovarian cancer susceptibility alleles. The 6 candidate ovarian cancer susceptibility genes are selected because they interact in the *BRCA1* and *BRCA2* pathways. These are *RAD51B*, *RAD51C*, *RAD51D*, *XRCC2*, *XRCC3* and *SLX4*. Cases are drawn from a several population based ovarian cancer case-control studies, and from two familial ovarian cancer registries. A proportion of the cases are previously analysed for mutations in *BRCA1* and *BRCA2*, in which case only individuals negative for pathogenic mutations in these genes are selected for further analysis. The study successfully sequences 1506 cases and 1130 healthy age-matched controls.

3.11 Target enrichment

Note: I prepared libraries for 3 sequencing lanes (1152 samples and 24 Access Array chips) at Great Ormond Street Molecular Genetics Laboratory. Maria Intermaggio and Andre Kim at USC prepared the remaining 4 lanes (1,536 samples and 32 Access Array chips).

Sequencing the coding region of 6 candidate genes from the human genome requires a method to select and enrich the regions of interest. The available methods for target enrichment are described in chapter 2. This study uses the Fluidigm Access Array System for both target enrichment and preparation of DNA sequencing libraries for the Illumina HiSeq2000 platform. The Fluidigm Access array is a PCR based method that generates sequencing ready, tagged amplicons that are 200 bp or less; thus the amplicons can be sequenced with 100 bp paired end reads. These tagged amplicons include both sequencing adaptors and individual indices for Illumina flow cell multiplexing.

Where exons are larger than 200 bp overlapping amplicons are designed to fully cover the coding region and flanking regions. Table 3.2 below describes the number of amplicons in the whole experiment and a breakdown of these showing numbers of

amplicons for each gene. It also gives a calculation of the percentage of the coding region covered for each gene by the amplicons. The genomic co-ordinates for each amplicon in each gene are given in the Appendix IV and an image, which maps the location of amplicons, is given in appendix V. The reference assembly used here is the Genome Reference Consortium Human genome build 37 (GRCh37/hg19) from February 2009. The National Centre for Biotechnology Information (NCBI) accession numbers for each gene are *RAD51B* (NM_133510.3), *RAD51C* (NM_058216.1), *RAD51D* (NM_002878.3), *XRCC2* (NM_005431.1), *XRCC3* (NM_005432.3) and *SLX4* (NM_032444.2).

Table 3.2 A breakdown of amplicons per gene

Gene	No. Amplicons	% of coding region included	5'UTR included (in bp)	3'UTR included (in bp)
RAD51B	18	100	99	93
RAD51C	17	100	52	60
RAD51D	15	100	31	67
XRCC2	11	100	63	49
XRCC3	13	100	48	80
SLX4	70	100	161	71
Total	144	100		

Table 3.2 A breakdown of amplicons per gene*. This table details the number of amplicons in the whole experiment, with the number for each gene and the percentage of the coding region included; the number of bases into 3' and 5' UTR for each gene is also given. Where there are more than 1 amplicon per exon amplicons are overlapping by more than the length of the primer sequence to ensure that all bases can be sequenced. Amplicons overlap into the flanking intronic regions by more bases than the length of primer sequences to include all the coding exons and splice sites*.

3.12 Library preparation

48 sequencing libraries are prepared on each Fluidigm Access Array chip. This includes the full coding region of all 6 candidate genes i.e. 144 PCR reactions are performed for 48 samples per chip. To reach this 3 PCR reactions are conducted in each well of the Access Array chip. PCR products are harvested and a second PCR is conducted to attach sample specific barcodes to each. Then each Access Array chip is then pooled to create a pool of 48 barcoded prepared libraries. Then 8 chips are pooled in equimolar quantities to form a pool of 384 individually barcoded libraries. The pools of 384 libraries are run in each flow cell lane on the Illumina HiSeq2000. The whole sequencing study is performed using 7 lanes of the HiSeq2000.

3.12.1 Quantitation of pools

As an initial quality control step, a proportion of the prepared libraries are quantified on the Agilent Bioanalyzer DNA 1000 chips using 1 µl of the library. This is performed to ensure that all samples have amplified prior to pooling. Images and results from these Agilent Bioanalyzer chips are included in Appendix VI. An example is given below in the gel-like image Figure 3.13.

Figure 3.13 Quantitation of Lane 4 (control samples)

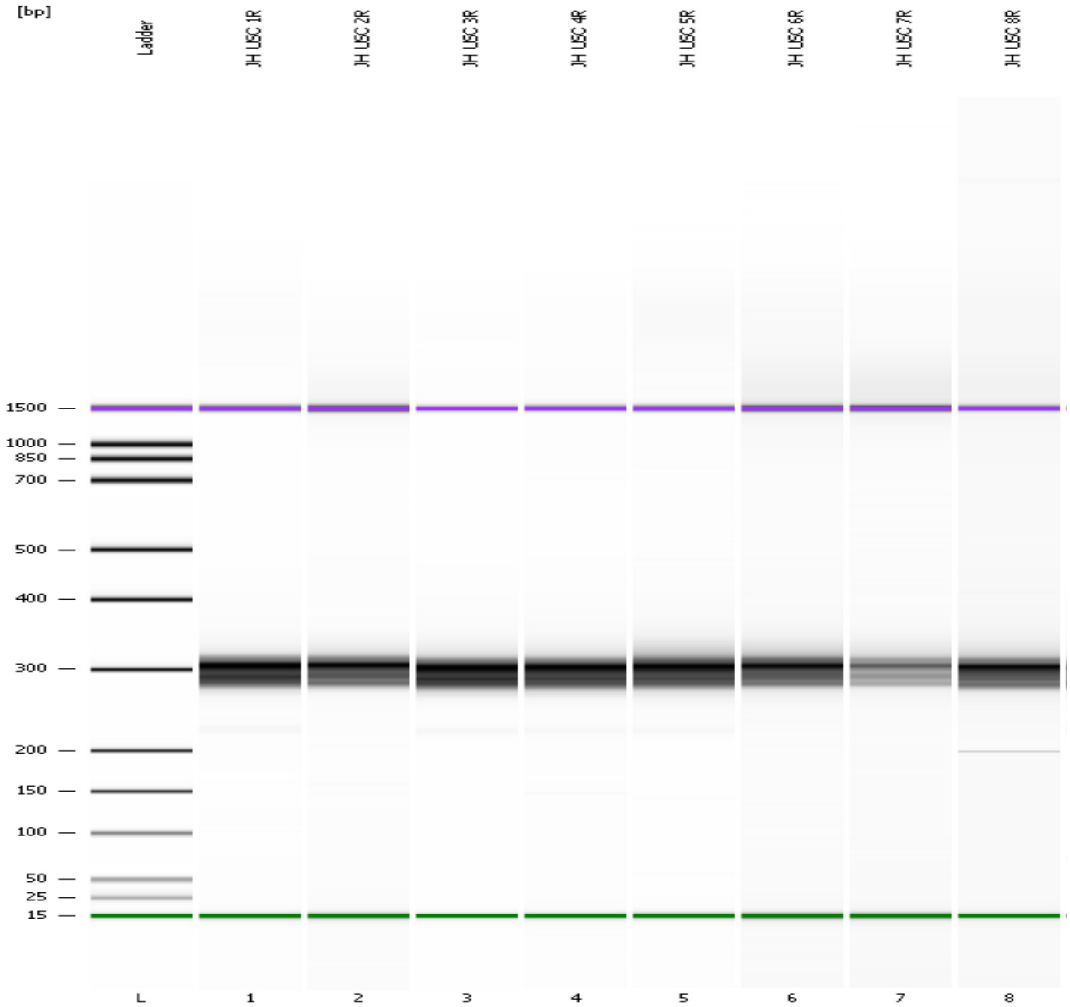


Figure 3.13 Quantitation for lane 4 (controls samples) using the Agilent Bioanalyzer*. A 300bp product is created that includes the target specific insert, the Illumina adapter sequences and the sample specific index sequences. This image shows the quantitation of 8 pools of Access Array chips each with 48 libraries. L = DNA Ladder, 1-8 = chips 1-8, the purple and green bands are the DNA markers.*

3.12.2 Normalisation of pools

For each Illumina flow cell lane pools of 8 chips are pooled in equimolar quantities and the final concentration is calculated to allow for dilution of pools to the correct concentration for the Illumina Flow cell.

Table 3.3 Normalisation table for pooling prepared libraries from 8 Fluidigm Access Array chips into one Illumina flow cell lane

Access Array Chip Number	Average Molarity of 10nM stock	Pool 1	Pooling Volume (µl)
Access Array Chip 1	47.30	Pool 1	21.1
Access Array Chip 2	49.60		20.2
Access Array chip 3	28.00		35.7
Access Array Chip 4	29.70		33.7
Access Array Chip 5	40.60		24.6
Access Array Chip 6	44.00		22.7
Access Array Chip 7	33.90		29.5
Access Array Chip 8	27.50		36.4

Table 3.3 Normalisation table for pooling prepared libraries. *This table shows an example of one pooled lane. The pools of 48 libraries, prepared using the Fluidigm Access Array, are pooled in equimolar quantities.*

Table 3.3 shows that the level of variation in molarity between pools for each chip is normalised to ensure that equal concentrations of samples are sequenced in each lane.

3.12.3. Final concentration

The final concentration required for the flow cell is 10nM in a total volume of 50 µl. The dilution factor is calculated from the initial molarity divided by the required final molarity. The volume of DNA to add is calculated by final volume divided by dilution factor. The volume water was added to reach 50 µl. Tables are in Appendices VII and VIII to show the final concentrations and dilutions for each lane.

3.13 Sequencing Quality Control

3.13.1 Phred scores (Q scores)

Quality control scores in next generation sequencing data are necessary for assessing the accuracy of data and for filtering sequencing artefacts. The Phred scoring system is used to assess base call accuracy. The Q score is a related to probability that a base

call is incorrect. Thus, Q = -10 log10 (P); this means that a Q score of 10 reveals that the probability that the base call is incorrect is 1 in 10 and the accuracy is 90%. A Q score of 30 means that the probability that the base call is incorrect is 1 in 1,000 and the accuracy is then 99.9%. Q30 is considered the standard minimum level of accuracy as this means that the majority of base calls are correct.

3.13.2 Read depth

Read depth, sometimes referred to as coverage, can be defined as the number of times a read is aligned to the reference. This is important in next generation sequencing to assess certainty in variant call accuracy. In the research setting a read depth of 30X (i.e. each of 30 reads are aligned to the reference sequence) is considered to be the benchmark for confident variant call accuracy. In the diagnostic clinic read depth of 50X is aimed for.

Table 3.4 Phred quality scores and read depth summary table

Lane	Mean Quality Score	Bases > Q30 (%)	Total number of reads in lane	Minimum Read Depth (X)	Maximum Read Depth (X)	Mean Read Depth per sample (X)
1	N/A	N/A	182,960,682	306	4248	1813
2	20.5	48.91	394,728,316	60	6829	3807
3	32.0	79.58	214,132,200	8	5437	2184
4	20.72	49.54	390,879,886	2	10443	3770
5	20.23	48.27	415,876,556	52	6284	4021
6	34.90	89.17	417,894,038	0	7214	4030
7	34.65	88.45	409,879,152	0	7525	3984

Table 3.4 Phred quality scores and read depth summary table*. This table summarises the quality scores, total number of reads in each lane, mean read depth for each lane and minimum and maximum read depth for each lane. X = number of reads, Q = Phred score, N/A=data not available*

The table above (Table 3.4) summarises the quality scores, mean read depth and the minimum and maximum read depth for each lane in the study. The table also includes Phred quality scores for each lane.

Phred quality scores were very good for lanes 3, 6 and 7 with 80% or more samples receiving a quality score of Q30 or more. None of the lanes are assigned quality scores below Q20, which means that there is a 1 in 100 chance that a base call is incorrect in these lanes. In other words there is a base call accuracy rate of 99% (Q20) and 99.9% (Q30).

In addition, figures 3.14 to 3.20 plot the individual sample read depths for each lane. This shows that, whilst in lanes 6 and 7 the minimum read depth is zero, very few samples failed in the whole study.

Figure 3.14 Graph plotting mean read depth per sample for lane 1

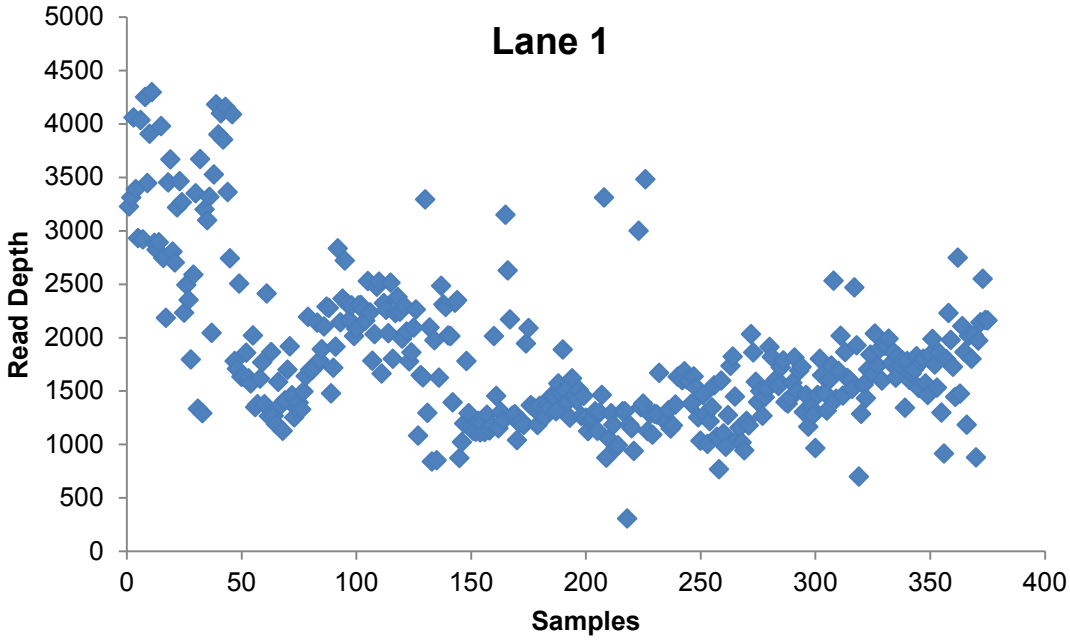


***Figure 3.14 Graph plotting mean read depth per sample for lane 1**. This graph shows the minimum mean read depth is 306 X and the maximum mean read depth is 4248 X*

Figure 3.15 Graph plotting mean read depth per sample for lane 2

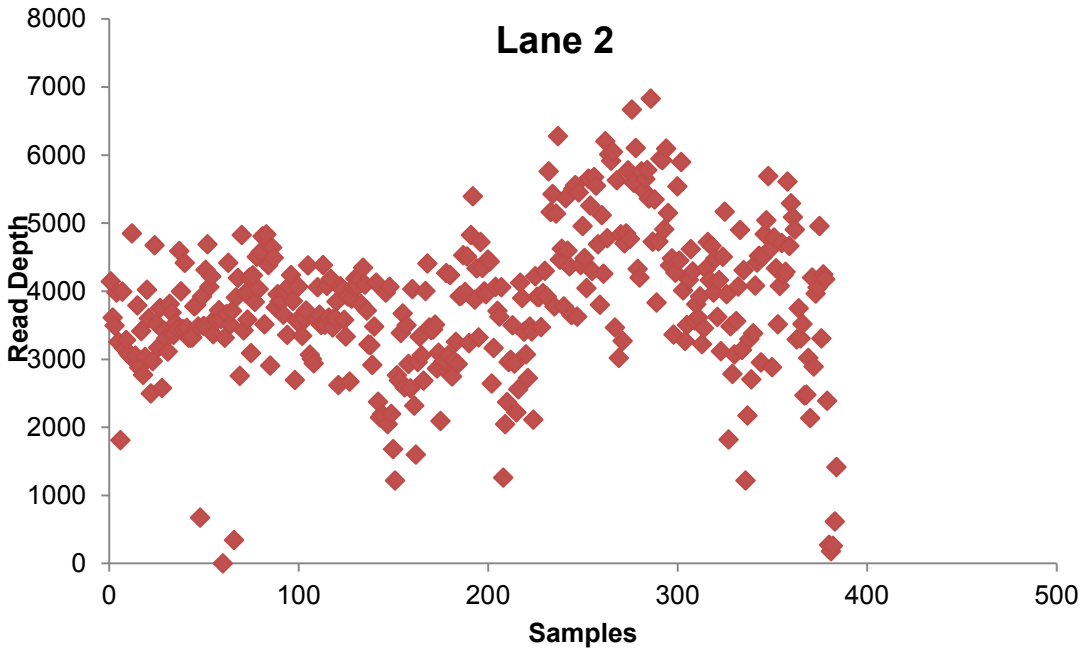


***Figure 3.15 Graph plotting mean read depth per sample for lane 2**. This graph shows the minimum mean read depth is 60 X and the maximum mean read depth is 6829 X*

Figure 3.16 Graph plotting mean read depth per sample for lane 3

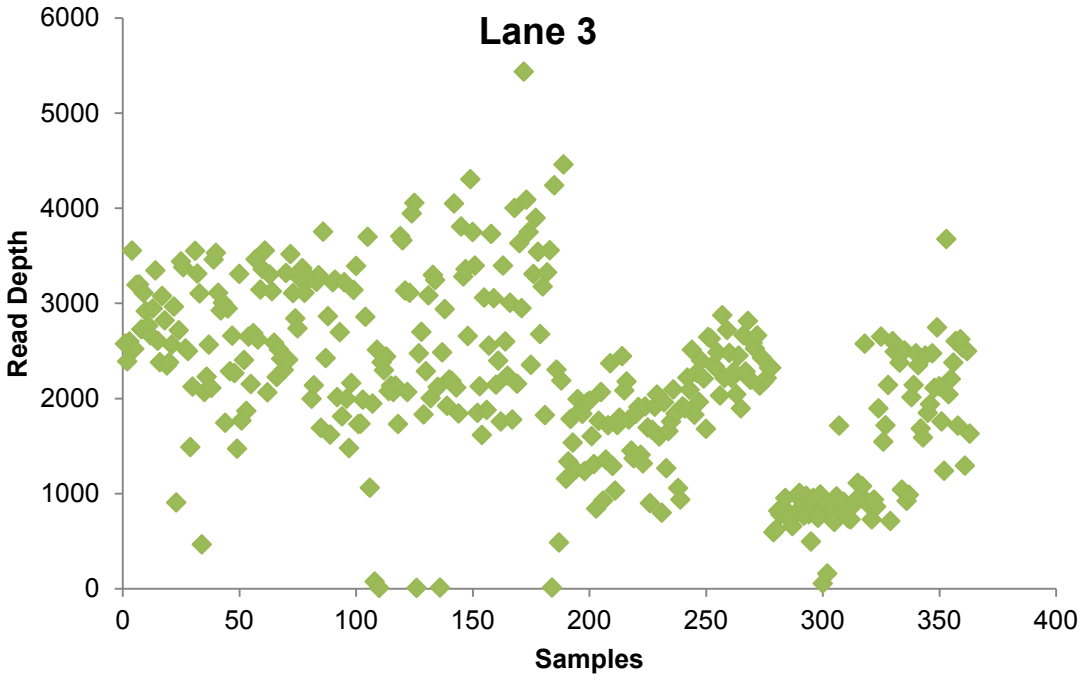


Figure 3.16 Graph plotting mean read depth per sample for lane 3*. This graph shows the minimum mean read depth is 8 X and the maximum mean read depth is 5437 X. 3 samples fall below read depth of 30 X.*

Figure 3.17 Graph plotting mean read depth per sample for lane 4

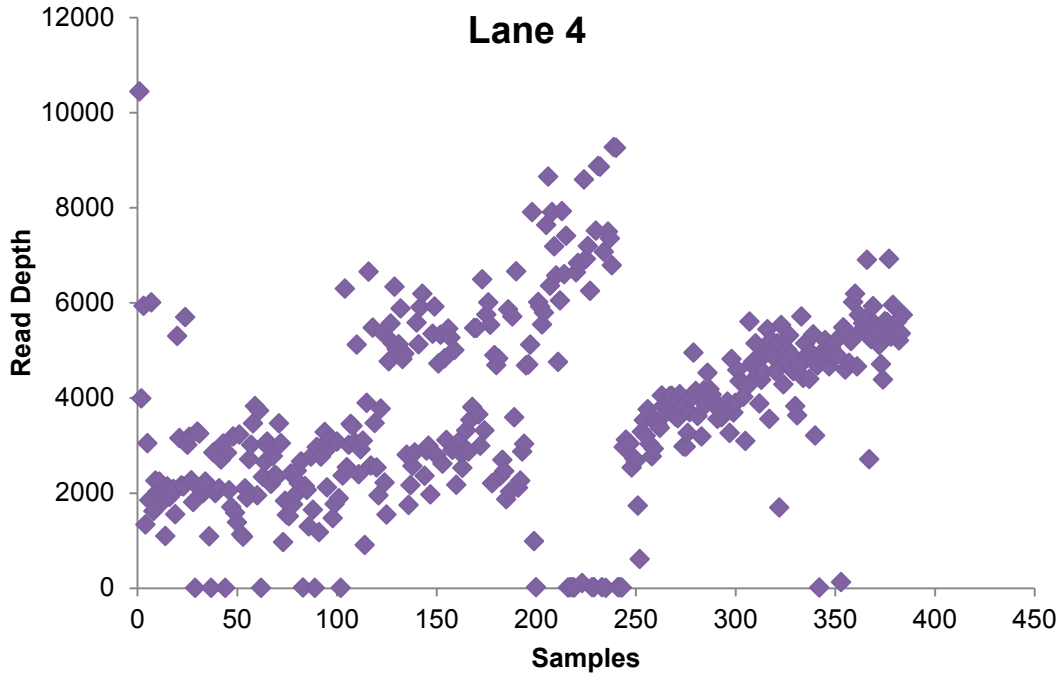


Figure 3.17 Graph plotting mean read depth per sample for lane 4*. This graph shows the minimum mean read depth is 2 X and the maximum mean read depth is 10443 X. 13 samples fall below read depth 30 X*

Figure 3.18 Graph plotting mean read depth per sample for lane 5

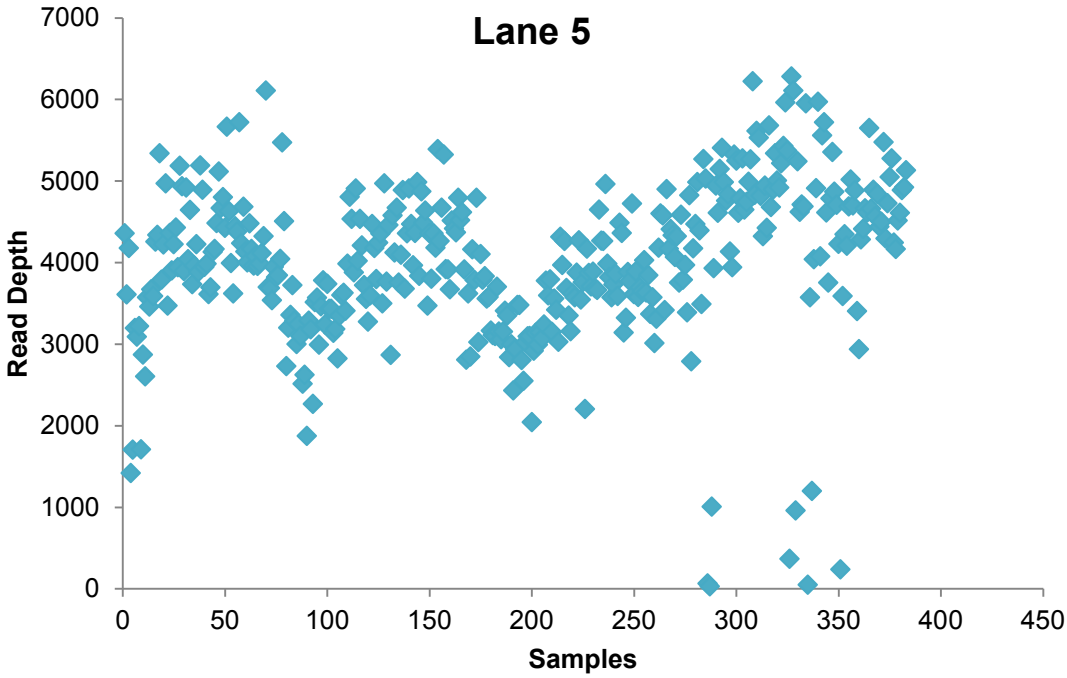


Figure 3.18 Graph plotting mean read depth per sample for lane 5*. This graph shows the minimum mean read depth is 52 X and the maximum mean read depth is 6284 X.*

Figure 3.19 Graph plotting mean read depth per sample for lane 6

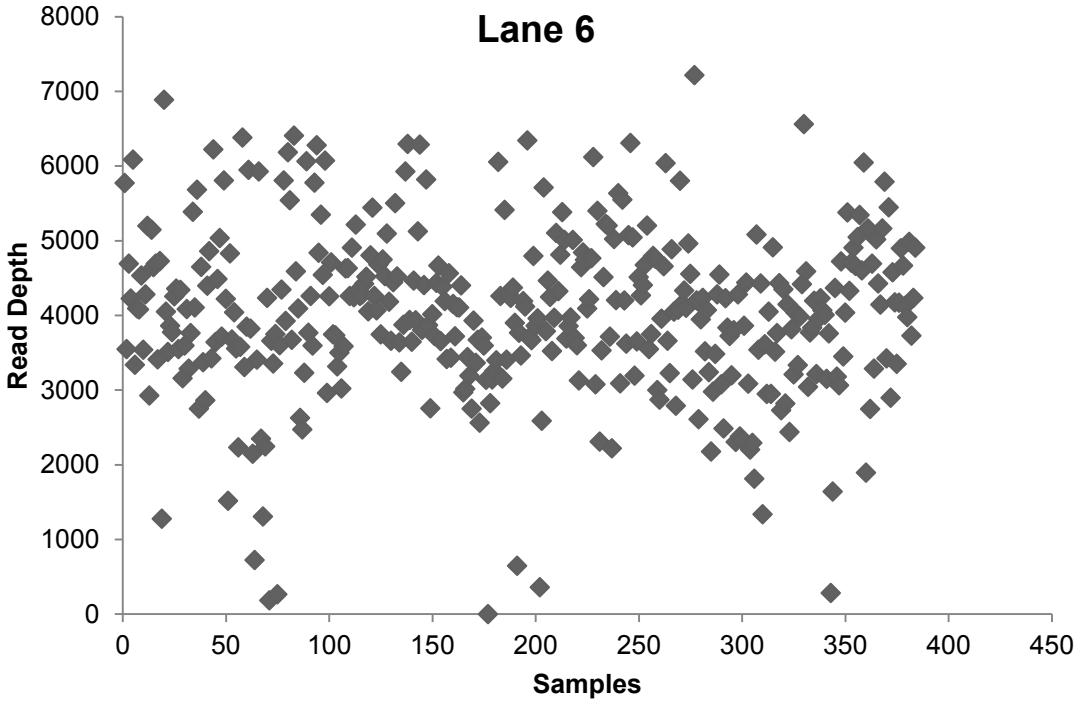


Figure 3.19 Graph plotting mean read depth per sample for lane 6*. This graph shows the minimum mean read depth is 71 X and the maximum mean read depth is 7214 X. The sample that shows read depth 0 X is the non-template control (NTC)*

Figure 3.20 Graph plotting mean read depth per sample for lane 7

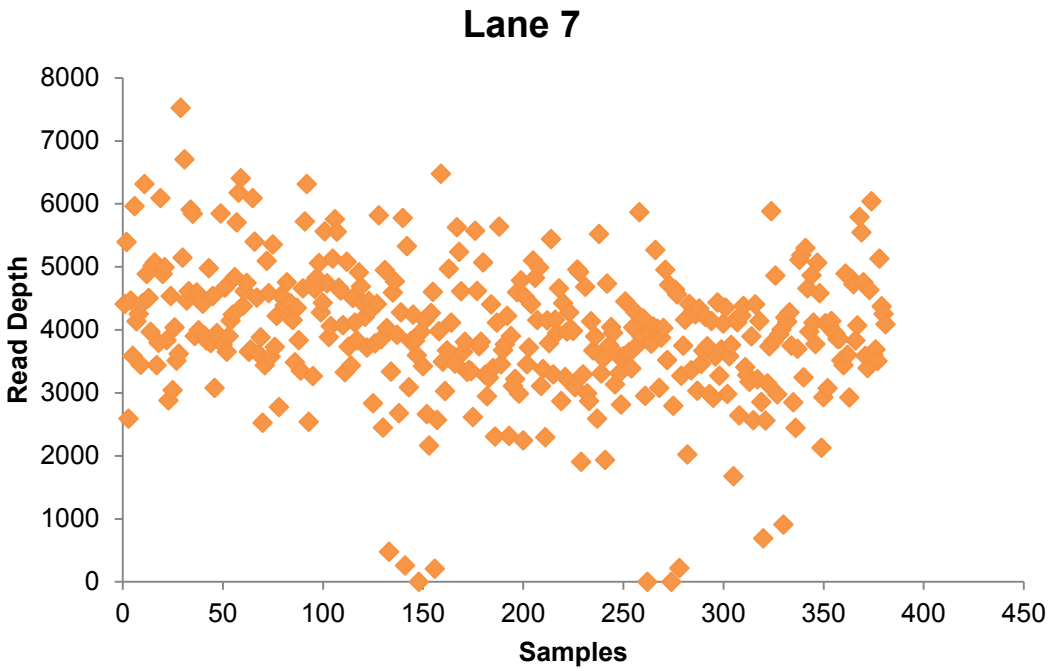


Figure 3.20 Graph plotting mean read depth per sample for lane 7*. This graph shows the minimum mean read depth is 206 X and the maximum mean read depth is 7525 X. The samples showing read depths 0 X are the non-template controls (NTC)*

In Figures 3.14 to 3.20 these graphs demonstrate that depth of coverage is generally even between samples and that most samples are sequenced at a read depth of that required for diagnostic mutation testing (i.e. 50X). Lane 4 is the least well performing lane as 13 samples failed to sequence and the coverage varies widely between samples. Lanes 6 and 7 are the best performing lanes as the majority of samples are sequenced with a mean read depth hovering around 4000 X with fewer outliers high or low. There are several reasons why these differences may be seen. It may be that there are differences in amplification between samples at the library preparation stage in lanes 3, 4 and 5 and in lane 6 and 7 the amplification may have been more even. The normalisation of samples is very important prior to amplification using the Access Array and this could be the reason for the unevenness in coverage. A second possibility is the normalisation and pooling of samples following library preparation and prior to loading on the C-bot for cluster generation.

Further analyses are conducted on read depth per amplicon. Each of the 144 amplicons are split into two, one for the forward read and one for the reverse read, which gives a total of 288 'amplicons' for each of 7 lanes. This means the total number of possible 'amplicons' is 2,016. For each 'amplicon' 50% of samples must be sequenced at a minimum of 30X coverage to be assessed as 'passed'. Table 3.5 gives

a breakdown of the number of failed and passed amplicons in each gene. This shows that overall 96.53% of amplicons are sequenced at a depth greater than 30X. The best performing gene is *XRCC2* with 99.35% sequenced at read depth greater than 30X and the poorest performing gene is *SLX4* with 95.2% sequenced at read depth greater than 30X.

Table 3.5 The number of passed and failed amplicons in each gene for the whole study

Gene	Passed	Failed	Total	% Passed
RAD51B	247	5	252	98.02
RAD51C	233	5	238	97.90
RAD51D	204	6	210	97.14
XRCC2	153	1	154	99.35
XRCC3	176	6	182	96.70
SLX4	933	47	980	95.20
Total	1946	70	2016	96.53

Table 3.5 The number of passed and failed amplicons in each gene for the whole study*. This table gives data for all possible 'amplicons' in the study. For each lane there are 288 possible 'amplicons' (forward and reverse reads) with a total of 2,016 across all 7 lanes. This table also shows the proportion of passing amplicons for each gene. For an 'amplicon' to pass >50% of samples must show a read depth of >30X. (Data supplied by Dr Honglin Song at Strangeways Research Laboratory, Cambridge UK).*

Further in depth analyses are conducted per sample; samples are considered 'failed' if greater than 80% of the sample has a read depth under 30X. Using these parameters, 51 samples are considered failed. For the remaining samples that are considered 'passed' the overall median depth of coverage is 2,264X (interquartile range is 1,502 – 3251).

For each gene the number of passed samples per gene is calculated to give a sensitivity figure for each gene and across all 6 genes. The Table 3.6 shows the sensitivity for the whole study, that is the proportion of samples in each gene sequenced at a read depth of >30X.

Table 3.6 Proportion of samples in each gene with a read depth >30X

Gene	Proportion >30X (%)
RAD51B	93
RAD51C	94
RAD51D	95
XRCC2	95
XRCC3	93
SLX4	93

Table 3.6 Proportion of samples with read depth >30X for each gene. *The overall mean sensitivity across all 6 genes is 94%. (Data supplied by Dr Honglin Song at Strangeways Research Laboratory, Cambridge UK).*

3.14 Genetic variant prevalence and characteristics

Genetic variants or changes in the DNA sequence may or may not cause disease (deleterious or pathogenic). Common polymorphisms are gene variants that occur in more than 1% of the population and rare variants occur at a rate less than 1% in the population. Common polymorphisms are generally considered to be non-pathogenic, thus in this study the common polymorphisms, as assessed either on dbSNP or through the frequency detected in the study, are filtered out and excluded as predicted neutral variants. Variants that are predicted to result in protein-truncation, i.e. those introducing a stop codon, frameshift insertions and deletions and variants at splice sites are considered predicted pathogenic variants for the purpose of this study. For definitive biological effect of novel variants detected functional assays would be required, therefore, these are referred to as predicted pathogenic variants.

3.14.1 Variant detection sensitivity and specificity

Variant detection sensitivity can be broadly defined as the proportion of variants detected and specificity as the proportion of false positives found in the data. Variants detected in NGS are validated here using Sanger sequencing to assess the false positives in the NGS approach. The positive controls spiked in for *RAD51C* is used to assess the accuracy of the NGS approach and to assess the accuracy of the multiplexing. Read filtering is conducted to filter out sequencing artefacts.

3.14.2 Blinded positive controls

6 blinded positive controls for *RAD51C* are spiked into these data as an assessment of variant detection sensitivity. 5 of these are accurately detected by NGS and the 6th control is not detected due to zero coverage for the region in which the variant resides. Figures 3.21 to 3.26 are a series of images representing the reads for each of the positive controls at the variant position.

Figure 3.21 Integrative Genome Viewer generated image for control No1 *RAD51C* c.-26C>T

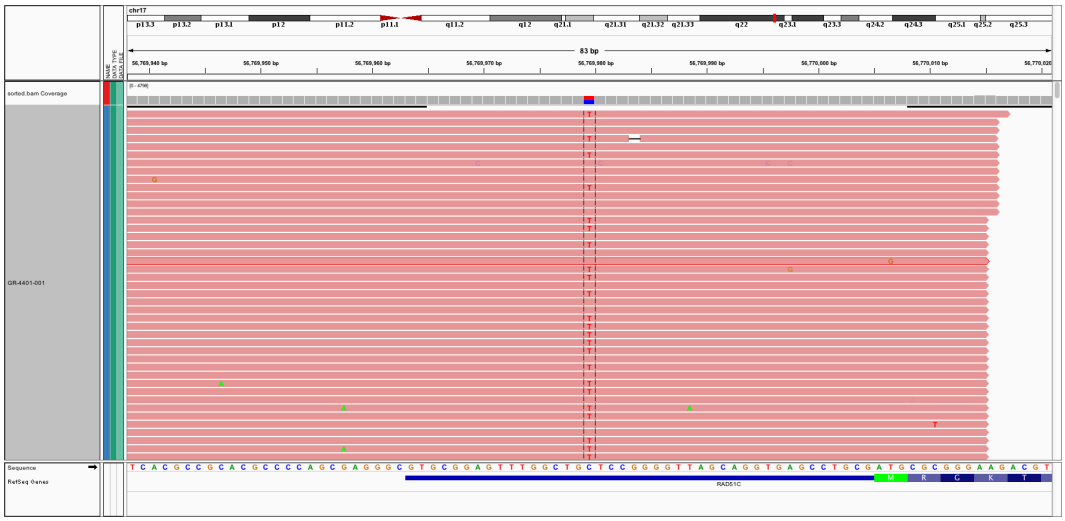


Figure 3.21 Integrative Genome Viewer generated image for control No1. *RAD51C c.-26C>T. The variant is clearly seen in the forward reads shaded pink and highlighted by parallel vertical dotted lines in the image*

Figure 3.22 Integrative Genome Viewer generated image for control No2 *RAD51C* c.374G>T

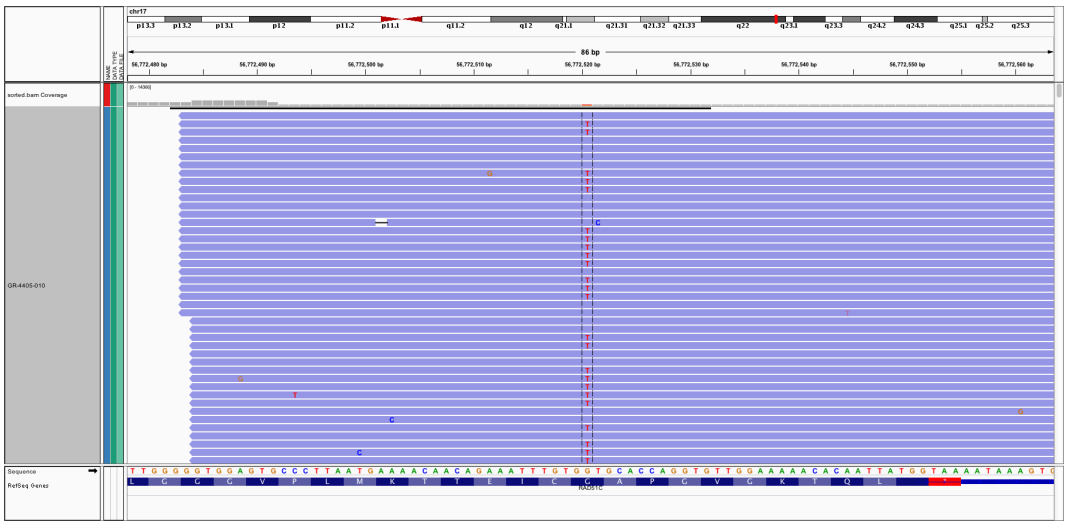


Figure 3.22 Integrative Genome Viewer generated image for control No2 RAD51C c.374G>T. *The variant is seen in the reverse reads shaded blue and highlighted by dotted parallel vertical lines*

Figure 3.23 Integrative Genome Viewer generated images control No3 *RAD51C* c.687C>T

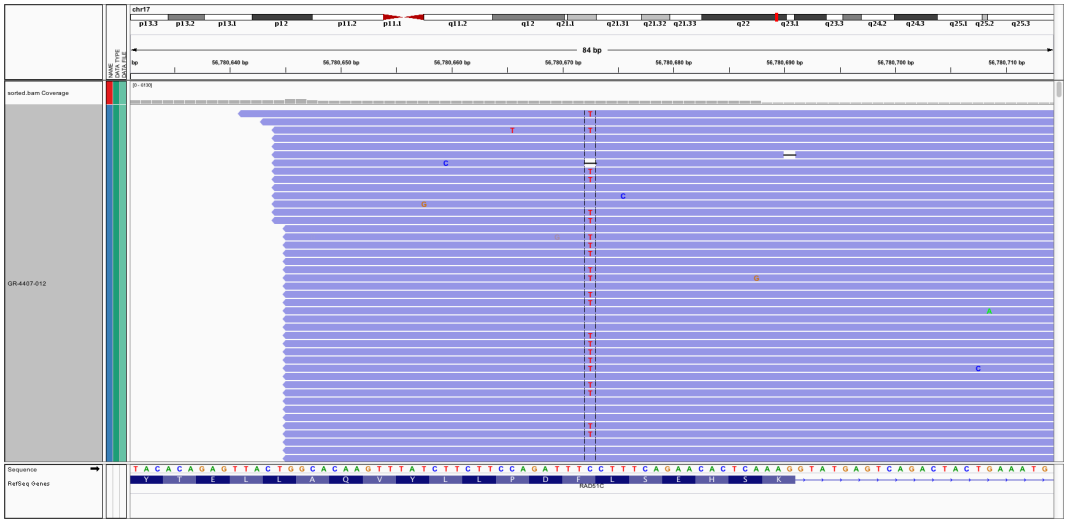


Figure 3.23 Integrative Genome Viewer generated images control No3 RAD51C c.687C>T. *The variant is seen in the reverse reads shaded blue and highlighted by dotted parallel vertical lines*

Figure 3.24 Integrative Genome Viewer generated images control No4 *RAD51C* c.IVS6(-19)T>C

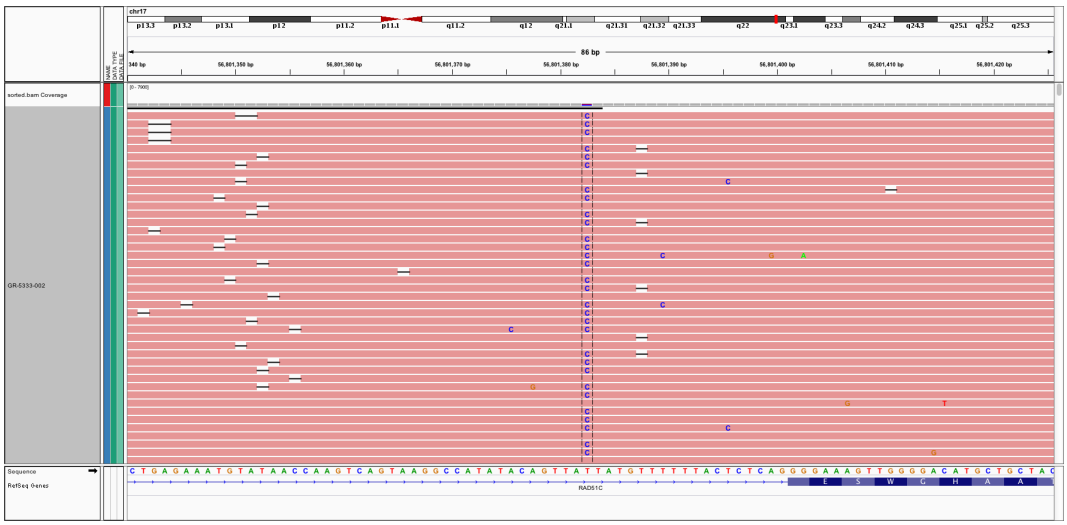


Figure 3.24 Integrative Genome Viewer generated images control No4 RAD51C c.IVS6(-19)T>C. *The variant is clearly seen in the forward reads shaded pink and highlighted by parallel vertical dotted lines in the image*

Figure 3.25 Integrative Genome Viewer generated images control No5 *RAD51C* c.790G>A

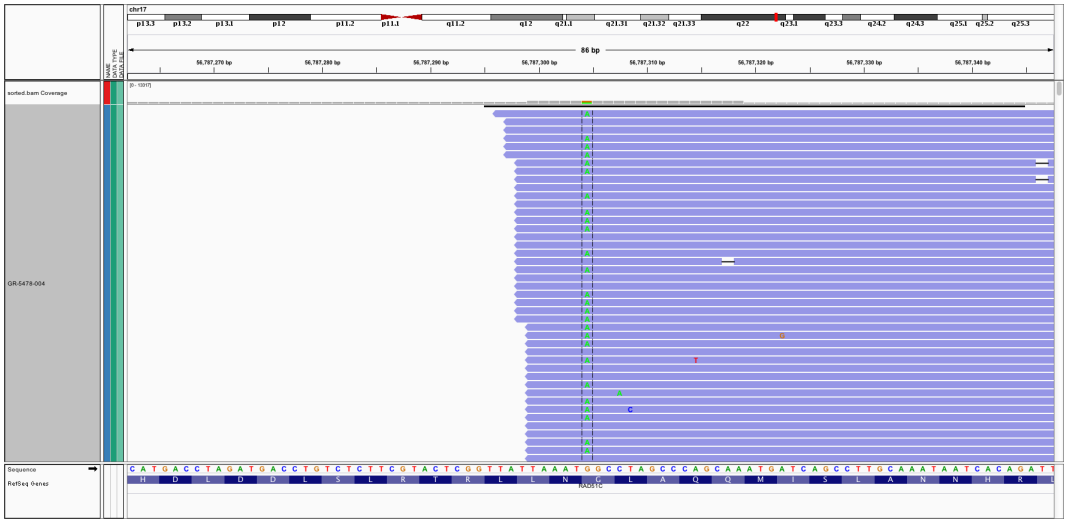


Figure 3.25 Integrative Genome Viewer generated images control No5 RAD51C c.790G>A. *The variant is seen in the reverse reads shaded blue and highlighted by dotted parallel vertical lines*

Figure 3.26 Integrative Genome Viewer generated image for control No6 *RAD51C* c.1097G>A

Figure 3.26 Integrative Genome Viewer generated image of control No.6. RAD51C c.1097G>A*. The image demonstrates that there is zero depth of coverage in the region highlighted with the parallel dotted lines in which the variant should be located. Therefore, this variant was missed due to no coverage.*

3.14.3 Filtering by read depth and alternate allele frequency

Filtering is performed to distinguish sequencing artefacts from real variants. The following parameters are used to filter out spurious variant calls. Variants detected are accepted if read depth is 15X or more and the alternate allele frequency is 40% or more; or if the read depth is 30X or more and the alternate allele frequency is 30% or more. Where variants are detected in regions that do not meet these criteria then variants are rejected. Table 3.7 gives the numbers of variants detected prior to filtering (unfiltered data) and the number of variants post filtering (filtered data)

Table 3.7 Filtering variants by lane

Lane	Unfiltered or Filtered	No.Variants
Lane 1	Unfiltered	1432
	Filtered	1354
Lane 2	Unfiltered	1482
	Filtered	1448
Lane 3	Unfiltered	1300
	Filtered	1167
Lane 4	Unfiltered	8684
	Filtered	2985
Lane 5	Unfiltered	1473
	Filtered	1295
Lane 6	Unfiltered	1539
	Filtered	1465
Lane 7	Unfiltered	1756
	Filtered	1616

Table 3.7 Filtering variants by lane. *Variants are filtered out according to read depth and alternate allele frequencies. Where read depth is greater than 15X alternate allele frequency must be greater than 40% and where read depth is greater than 30X alternate allele frequency must be greater than 30%.*

3.14.4 Filtering out silent variants

Following filtering by read depth and alternate allele frequency all synonymous single nucleotide variants (S-SNV) that do not result in a change in amino acid are removed in the next stage of the analysis. This leaves only the non-synonymous single nucleotide variants (NS-SNV) that result in a change in amino acid and the insertions and deletions (INDELs). Table 3.8 details the total number of remaining variants that lead to a change in amino acid, some of which are the same variant in different samples.

Table 3.8 Number of remaining variants following removal of silent variants

Lane	No. remaining variants post removal of silent variants
Lane 1	802
Lane 2	855
Lane 3	692
Lane 4	2168
Lane 5	846
Lane 6	887
Lane 7	1048

***Table 3.8 Number of remaining variants following removal of silent variants**. All synonymous single nucleotide variants are removed in this stage of the analysis. Note that some of these variants are the same change in different samples*

3.14.5 Summary of genetic variants detected according to variant type.

Following filtering of S-SNVs, variants are subdivided into non-synonymous single nucleotide variants (NS-SNVs), frameshift insertions and deletions, nonsense variants and splice site variants. The table 3.9 summarises all variants detected in the study by gene giving number of different variants of each type.

Table 3.9 Summary of genetic variants detected according to variant type

Gene	Nonsense	Frameshift	Splice site	NS-SNV
RAD51B	1	0	0	13
RAD51C	4	2	2	18
RAD51D	3	1	0	16
XRCC2	0	1	0	11
XRCC3	0	0	1	18
SLX4	2	1	0	27
Total	10	5	3	103

***Table 3.9 Summary of genetic variants detected according to variant type**. This table gives detail on the total number of different variants detected by variant type for the whole study. (NS-SNV = non-synonymous single nucleotide variant)*

3.15 Predicted deleterious variants detected in the 6 genes.

Nonsense, frameshift and splice site variants are assumed to be predicted deleterious variants as these are anticipated to result in protein truncation. The NS-SNVs are not assumed to be deleterious and the functional effects of these are predicted using two software programs PROVEAN and PolyPhen-2.

3.15.1 Predicted protein-truncating variants

Table 3.10 Predicted protein-truncating variants

Gene	Function	Variant cDNA	Exon	No. cases or controls	Novel or known	Sanger sequencing validation Yes/No
RAD51B	Nonsense	c.489T>G	6	1 case	Novel	Yes
RAD51C	FS Del	c.497delT	3	1 case	Novel	Yes
RAD51C	FS Del	c.651_652del	4	1 case	Novel	Yes
RAD51C	Nonsense	c.577C>T	4	2 cases	Known	Yes
RAD51C	Nonsense	c.955C>T	7	2 cases	Novel	Yes
RAD51C	Splicing	c.706-2A>G	5	2 cases	Known	Yes
RAD51C	Splicing	c.905-2delAG	7	1 case	Known	Yes
RAD51C	Nonsense	c.1005C>A	8	1 case	Novel	No
RAD51C	Nonsense	c.312T>A	2	1 case	Novel	Unknown
RAD51D	FS Del	c.565_568del	6	2 cases	Novel	Yes
RAD51D	Nonsense	c.478C>T	5	1 case	Novel	Yes
RAD51D	Nonsense	c.620C>A	7	1 case	Novel	Yes
RAD51D	Nonsense	c.898C>T	9	1 control	Novel	Yes
XRCC2	FS Del	c.96delT	2	2 cases	Novel	Yes
XRCC3	Splicing	c.194-2A>G	7	1 case	Novel	Yes
SLX4	FS Del	c.2497_2498del	12	1 Case	Novel	Yes
SLX4	Nonsense	c.4386_4387insAGGATGAACGAGGCCGC	12	2 Cases	Novel	Yes
SLX4	Nonsense	c.1976C>A	9	1 case	Novel	No
SLX4	Non FS Del	c.3919_3921del	12	1 control	Novel	Yes

Table 3.10 Predicted protein-truncating variants. *This table gives details of all predicted protein-truncating variants detected in all 6 genes in the whole study. Variants are named according to the cDNA sequence of each gene. There are 19 different variants of which Sanger sequencing validates 16. 16 variants are novel with 3 previously reported in literature. 20 cases have validated predicted protein-truncating variants; 1 case has a predicted protein-truncating variant with unknown validation status; 2 cases with predicted protein-truncating variants are not validated by Sanger sequencing and 1 control has a predicted protein-truncating variant and 1 control has a non-frameshift deletion.*

The predicted protein-truncating variants for the whole study are described in the table 3.10. This table gives details of type of variants detected in each gene and if they are novel or known. Table 3.10 also gives the number of variants in cases and controls. There are 19 different variants in total with 16 of those validated by Sanger sequencing. 3 variants are previously reported in literature: the nonsense variant in *RAD51C* c.577C>T and the splice site variant c.905-2delAG have been published by Coulet et al (2012) and the splice site variant in *RAD51C* 706-2A>G is published by Loveday et al (2012). All other variants are previously unknown.

3.15.2 Predicted deleterious non-synonymous single nucleotide variants

In addition to predicted protein truncating variants in the 6 genes under study, several other genetic variants, including putative functional missense alterations and known polymorphisms are identified in these data. Table 3.11 gives the results from two functional prediction programs (PROVEAN and PolyPhen-2).

Table 3.11 Results from PROVEAN and PolyPhen-2 software functional prediction programs for each of the non-synonymous single nucleotide variants

Gene	Variant	PROVEAN	PolyPhen-2
RAD51B	L172W	Neutral	Possibly damaging
	K243R	Neutral	Possibly damaging
	V207L	Neutral	Benign
	C185G	Deleterious	Possibly damaging
	Y180C	Neutral	Benign
	R217G	Deleterious	Probably damaging
	A89T	Neutral	Benign
	E340Q	Neutral	Benign
	M120T	Neutral	Benign
	A295V	Deleterious	Probably damaging
	D142G	Deleterious	Probably damaging
	S250A	Neutral	Benign
	V343A	Neutral	Possibly damaging
RAD51C	G264S	Deleterious	Benign
	T287A	Deleterious	Probably damaging
	G125V	Deleterious	Probably damaging
	A126T	Neutral	Benign
	L226P	Deleterious	Probably damaging
	A354V	Neutral	Benign
	D318N	Neutral	Benign
	V169G	Deleterious	Possibly damaging
	K84N	Neutral	Benign
	L27P	Deleterious	Probably damaging
	L262V	Neutral	Possibly damaging
	T174S	Neutral	Benign
	E67K	Neutral	Benign
	L91F	Deleterious	Probably damaging
	P43S	Deleterious	Possibly damaging
	Q222K	Neutral	Possibly damaging
	Q268H	Neutral	Possibly damaging
	Q11R	Neutral	Possibly damaging
RAD51D	R165Q	Neutral	Benign
	A210V	Deleterious	Probably damaging
	C119R	Neutral	Possibly damaging

	C117S	Deleterious	Probably damaging
	C9S	Deleterious	Possibly damaging
	H23Y	Neutral	Benign
	Q115R	Deleterious	Probably damaging
	R165W	Deleterious	Probably damaging
	R239W	Deleterious	Probably damaging
	T313A	Neutral	Benign
	L164P	Deleterious	Probably damaging
	M308V	Neutral	Benign
	G44A	Neutral	Benign
	I251M	Neutral	Benign
	E233G	Deleterious	Probably damaging
	V56G	Deleterious	Possibly damaging
XRCC2	R188H	Neutral	Benign
	D36N	Neutral	Benign
	V118E	Deleterious	Benign
	E207G	Neutral	Benign
	R214Q	Neutral	Benign
	K229R	Neutral	Benign
	T94R	Neutral	Possibly damaging
	V39M	Neutral	Possibly damaging
	F240L	Neutral	Benign
	F32V	Deleterious	Probably damaging
	L90F	Deleterious	Benign
XRCC3	T241M	Neutral	Possibly damaging
	A262T	Neutral	Benign
	V194M	Neutral	Possibly damaging
	R108C	Neutral	Benign
	R162C	Deleterious	Probably damaging
	K17N	Neutral	Possibly damaging
	R94H	Neutral	Benign
	R302H	Neutral	Benign
	K22R	Neutral	Benign
	S110L	Deleterious	Probably damaging
	V124M	Deleterious	Probably damaging
	V165I	Neutral	Benign
	R58W	Deleterious	Probably damaging
	P230L	Neutral	Benign
	P230S	Neutral	Benign
	R313W	Deleterious	Probably damaging
	R231K	Neutral	Probably damaging
	M263V	Neutral	Benign
SLX4	A424V	Deleterious	Probably damaging
	A535V	Deleterious	Probably damaging

	D1802N	Deleterious	Probably damaging
	E787K	Deleterious	Possibly damaging
	K354N	Deleterious	Probably damaging
	K458E	Deleterious	Probably damaging
	L472S	Deleterious	Probably damaging
	L497S	Deleterious	Probably damaging
	P1122L	Deleterious	Benign
	P1381S	Deleterious	Possibly damaging
	P1624A	Deleterious	Benign
	P504R	Deleterious	Probably damaging
	Q273H	Deleterious	Probably damaging
	Q355L	Deleterious	Probably damaging
	R1341I	Deleterious	Possibly damaging
	R1550L	Deleterious	Possibly damaging
	R1550W	Deleterious	Benign
	R1814C	Deleterious	Probably damaging
	R204C	Deleterious	Probably damaging
	R278W	Deleterious	Probably damaging
	R811T	Deleterious	Probably damaging
	S1123Y	Deleterious	Possibly damaging
	S1271F	Deleterious	Probably damaging
	S1492Y	Deleterious	Probably damaging
	S498F	Deleterious	Probably damaging
	T757I	Deleterious	Probably damaging
	V362A	Deleterious	Probably damaging

Table 3.11 Putative functional missense mutations in the 6 genes under study: *two software programs PROVEAN and PolyPhen-2 are used to predict the putative functional effects of missense variants. Results suggest that there is a high degree of concordance between deleterious prediction by PROVEAN and probably damaging by PolyPhen-2. Differences are seen in the additional level for PolyPhen-2 the 'Possibly damaging' subset.*

A final set of 56 predicted functional missense variants for all 6 genes under investigation is compiled in Table 3.12. Sanger sequencing validation results are available for 5 of these variants; this is due to the high costs that would have incurred in full validation as there are very large numbers of samples involved. 4 out of 5 Sanger sequenced variants are confirmed and these are highlighted Table 3.12 (green=validated and pink=not validated). Table 3.12 also gives details on dbSNP listing with rs SNP numbers where known and minor allele frequencies (MAF) if known. MAF is taken from 1000 Genomes data.

Table 3.12 Predicted functional missense variants

Gene	Variant (Protein)	Variant (gDNA)	Position (Exon No.)	Frequency (%)		rs dbSNP	MAF	Source if known	Sanger Validation Y/N/Unknown
				Cases	Controls				
RAD51B	C185G	c.553T>G	6	0.07	0	Novel			Unknown
RAD51B	R217G	c.649A>G	7	0.07	0	Novel			Unknown
RAD51B	A295V	c.884C>T	9	0	0.09	Novel			Unknown
RAD51B	D142G	c.425A>G	5	0	0.09	Novel			Unknown
RAD51C	G264S	c.790G>A	5	0.60	0.62	rs147241704	N/A	dbSNP	Unknown
RAD51C	T287A	c.859A>G	6	1.39	0.97	rs28363317	G=0.006	1000 Genomes	Unknown
RAD51C	G125V	c.374G>T	2	0.07	0.27	rs267606998	N/A	Meindl et al (2010)	Unknown
RAD51C	L226P	c.677T>C	4	0.07	0	Novel			Yes
RAD51C	V169G	c.506T>G	3	0.13	0	Novel			No
RAD51C	L27P	c.80T>C	1	0.07	0	Novel			Unknown
RAD51C	L91F	c.271C>T	2	0.07	0	Novel			Unknown
RAD51C	P43S	c.1271C>T	1	0.07	0	Novel			Unknown
RAD51D	Q115R	c.344A>G	4	0.07	0	Novel			Unknown
RAD51D	C117S	c.349T>A	5	0.07	0	Novel			Unknown
RAD51D	C9S	c.26G>C	1	0.20	0	rs1408257595			Unknown
RAD51D	A210V	c.629C>T	7	0.13	0	Novel			Unknown
RAD51D	R165W	c.493C>T	6	0	0.18	Novel			Unknown
RAD51D	R239W	c.715C>T	10	0.07	0	Novel			Unknown
RAD51D	L164P	c.491T>C	6	0.07	0	Novel			Yes
RAD51D	E233G	c.698A>G	8	2.99	1.24	rs28363284	C=0.006	1000 Genomes	Unknown
RAD51D	V56G	c.167T>G	3	0.13	0	Novel			Unknown
XRCC2	F32V	c.94T>G	2	0.13	0	Novel			Yes
XRCC2	V118E	c.353T>C	3	0	0.09	rs185815454	G=0.001	1000 Genomes	Unknown
XRCC2	L90F	c.286C>T	3	0.07	0	Novel			Unknown
XRCC3	R162C	c.484C>T	7	0.07	0	Novel			Unknown
XRCC3	S110L	c.329C>T	6	0.07	0	Novel			Yes
XRCC3	R58W	c.172C>T	5	0	0.09	rs143410843	NA	Unknown	Unknown
XRCC3	V124M	c.370G>A	6	0.07	0	Novel			Unknown
XRCC3	R313W	c.937C>T	10	0.07	0	Novel			Unknown
SLX4	A424V	c.1271C>T	6	0.07	0	Novel			Unknown
SLX4	A535V	c.1604C>T	7	0.07	0	Novel			Unknown
SLX4	D1802N	c.5404G>A	15	0.07	0.09	Novel			Unknown
SLX4	E787K	c.2359G>A	12	0.27	0.18	rs140600202	NA	Unknown	Unknown
SLX4	K354N	c.1062G>T	5	0	0.09	Novel			Unknown
SLX4	K458E	c.1372A>G	7	0.07	0	rs149126845	0.000	1000 Genomes	Unknown
SLX4	L472S	c.1415T>C	7	0.13	0	Novel			Unknown
SLX4	L497S	c.1409T>C	7	0.07	0	Novel			Unknown
SLX4	P1122L	c.3365C>T	12	14.94	12.21	rs714181	A=0.197	1000 Genomes	Unknown
SLX4	P1381S	c.4141C>T	12	0	0.09	Novel			Unknown
SLX4	P1624A	c.4870C>G	14	0	0.18	Novel			Unknown
SLX4	P504R	c.1511C>G	7	0	0.09	Novel			Unknown
SLX4	Q273H	c.819G>C	4	0.07	0	Novel			Unknown
SLX4	Q355L	c.1064A>T	5	0	0.09	Novel			Unknown

SLX4	P1341I	c.4022G>T	12	0.07	0	Novel			Unknown
SLX4	R1550L	c.4649G>T	13	0	0.62	Novel			Unknown
SLX4	R1550W	c.4648C>T	4	0.66	0.71	rs77021998	A=0.002	1000 Genomes	Unknown
SLX4	R1814C	c.5440C>T	15	0	0.09	Novel			Unknown
SLX4	R204C	c.610C>T	3	12.22	11.33	Novel			Unknown
SLX4	R278W	c.832C>T	4	0.07	0	Novel			Unknown
SLX4	R811T	c.2432G>C	12	0.07	0	Novel			Unknown
SLX4	S1123Y	c.3368C>A	12	0.20	0	Novel			Unknown
SLX4	S1271F	c.3812C>T	12	6.44	5.04	Novel			Unknown
SLX4	S149Y	c.4475C>A	12	0.07	0	Novel			Unknown
SLX4	S498F	c.1493C>T	7	0.07	0	Novel			Unknown
SLX4	T757I	c.2270C>T	11	0.07	0	Novel			Unknown
SLX4	V362A	c.1085T>C	5	0.27	0.09	Novel			Unknown

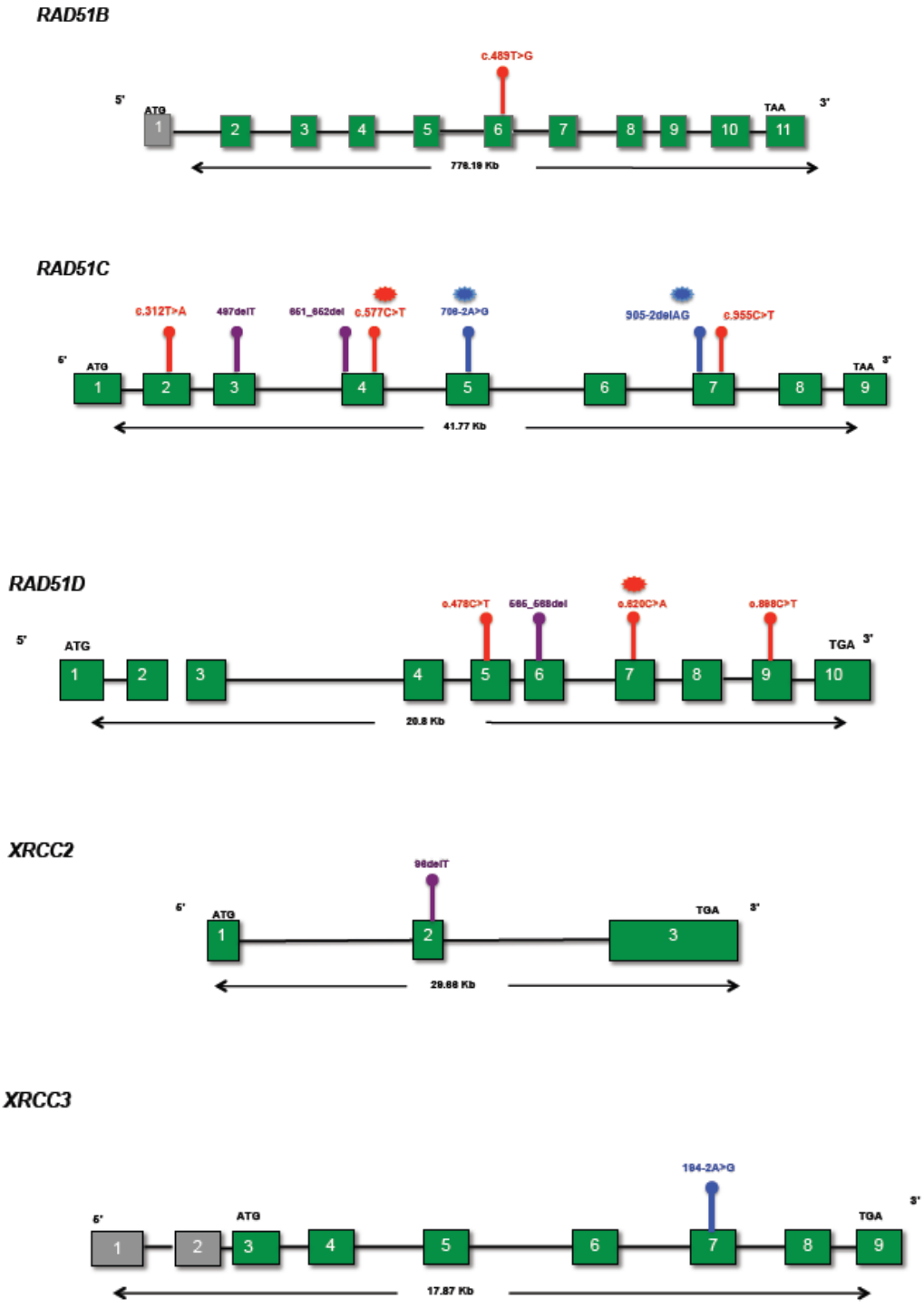
Table 3.12 Predicted functional missense variants. *This table describes all predicted functional missense variants detected in all 6 genes with frequency of each variant in cases and controls. The protein names are included as these are used in the functional prediction programs. Sanger sequencing validation is performed on a subset of 5 variants due to the large number of samples involved. 4 out of 5 variants are confirmed via Sanger sequencing. rsSNP number is included where these are known. MAF (Minor Allele Frequency) is given for the second most commonly occurring allele according to dbSNP in order to differentiate rare variants from common polymorphisms. MAF is based on frequency in 1000 genomes data and MAF G=0.008 means that allele G occurs at a frequency of 0.8% in the 1094 samples sequenced in 1000 Genomes data in 2011.*

3.16 Summary of variant prevalence and characteristics including position of variants in each gene

NGS data identifies 23 ovarian cancer cases and 1 control as having a predicted protein-truncating variant in one of the 6 genes. Of these Sanger sequencing validation confirms 20 cases and 1 control with predicted protein-truncating variants. In addition, 1 control sample is detected with a non-frameshift deletion. These variants include, 1 nonsense variant in *RAD51B,* 2 nonsense and 2 splice site variants in *RAD51C*, 1 frameshift and 3 nonsense variants in *RAD51D*, 1 frameshift deletion in *XRCC2,* 1 splice site variant in *XRCC3* and 1 frameshift deletion, 1 nonsense and 1 non-frameshift deletion in *SLX4*. In all 6 genes in the study 1.33% of cases have a predicted protein-truncating variant and just 0.09% of control samples have a predicted protein-truncating variant in one of the 6 genes.

The figure 3.27 is an image demonstrating the position of variants in each gene. This illustrates that variants have been detected throughout the coding region of each gene. This figure also shows which variants are novel and which were previously identified.

Figure 3.27 Images of each gene with position of predicted protein truncating variants



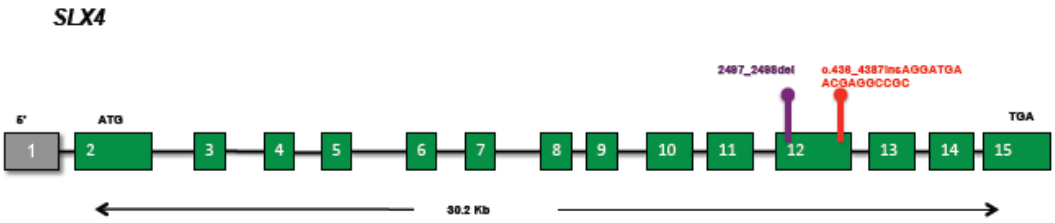


***Figure 3.27 Images of each gene with position of predicted protein truncating variants**. The variant in RAD51B is in exon 6, RAD51C has the most variants and these are positioned in exons 2, 3, 4, 5, 6 and 7, RAD51D has the second most frequently occurring variants and these are located in exons 5, 6, 7 and 9 only. XRCC2 and XRCC3 have only 1 variant in each gene. SLX4 has 2 variants and both of these are located in exon 12, the largest exon in the gene. Red = Nonsense. Blue = splicing. Purple = frameshift. The previously reported mutations are indicated with a star*

3.17 Summary of Sanger sequencing validation

Sanger sequencing validation confirms 20 cases with predicted protein-truncating variants and 1 predicted protein-truncating variant in 1 control sample. Figures 3.28 to 3.33 show images of examples of comparisons of Sanger sequencing data and NGS data for the same variant and sample. Integrative Genome Viewer (IGV) images of NGS data were generated and compared to the Sanger sequencing trace images.

Figure 3.28 Sanger sequencing trace and NGS IGV generated image for variant *RAD51B* c.489T>G

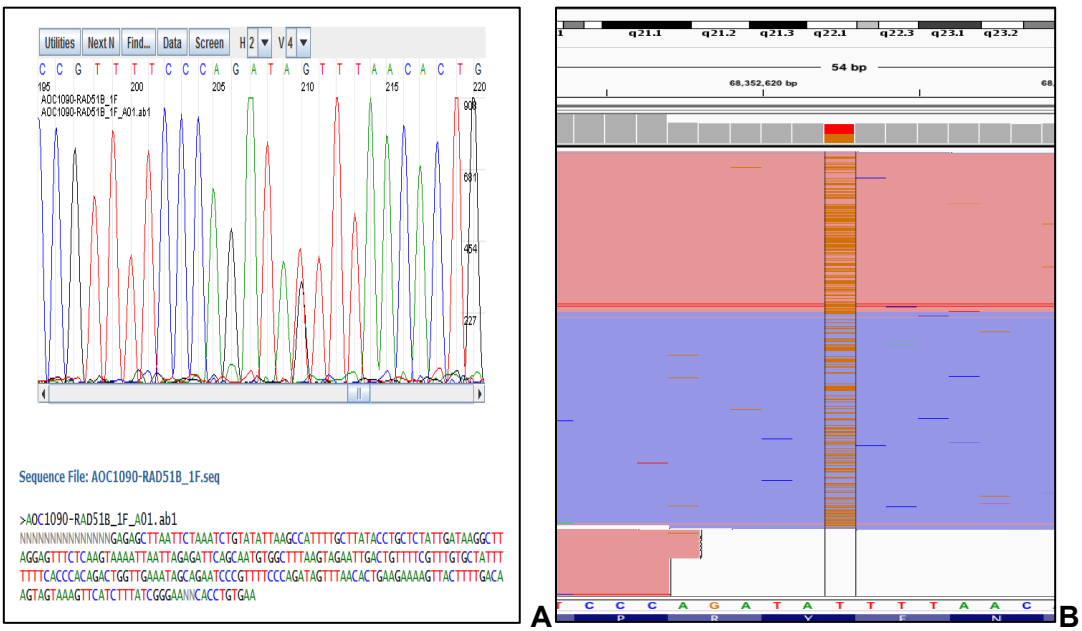


***Figure 3.28 Sanger sequencing trace and NGS IGV generated image for variant RAD51B c.489T>G** This figure shows images of the same RAD51B (c.489T>G) nonsense mutation, A. Sanger sequencing trace and B. NGS IGV generated image. The NGS image (B) shows reads 'squished' demonstrating that the mutation is detected in both directions (red = forward, blue = reverse).*

Figure 3.29 Sanger sequencing trace and NGS IGV generated image for variant *RAD51C* c.577C>T

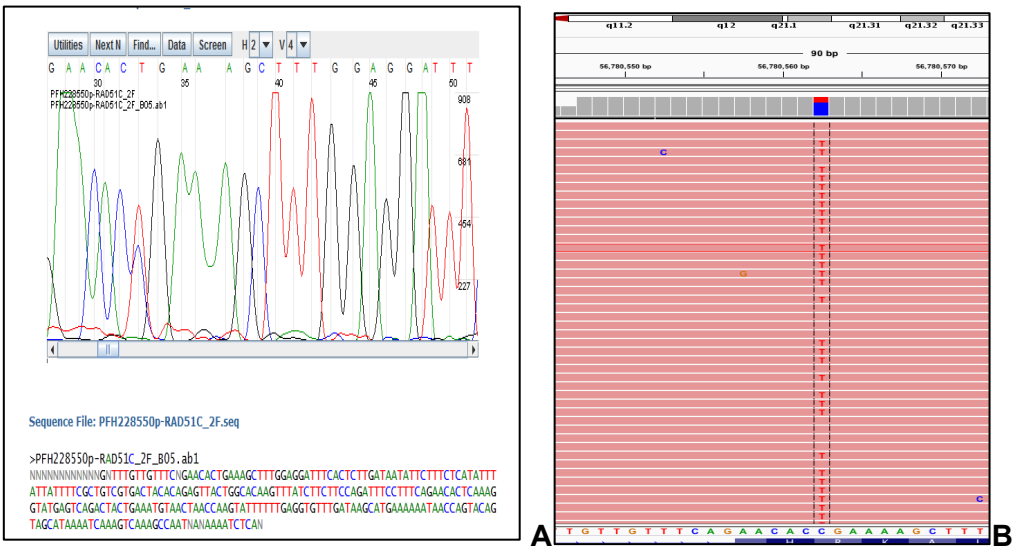


Figure 3.29 Sanger sequencing trace and NGS IGV generated image for variant RAD51C c.577C>T*. This figure shows images of the same RAD51C (c.577C>T) nonsense mutation, A. Sanger sequencing trace and B. NGS IGV generated image. The NGS image clearly shows the variant is heterozygous, with approximately half of the reads showing each allele.*

Figure 3.30 Sanger sequencing trace and NGS IGV generated image for variant *RAD51C* c.905-2delAG

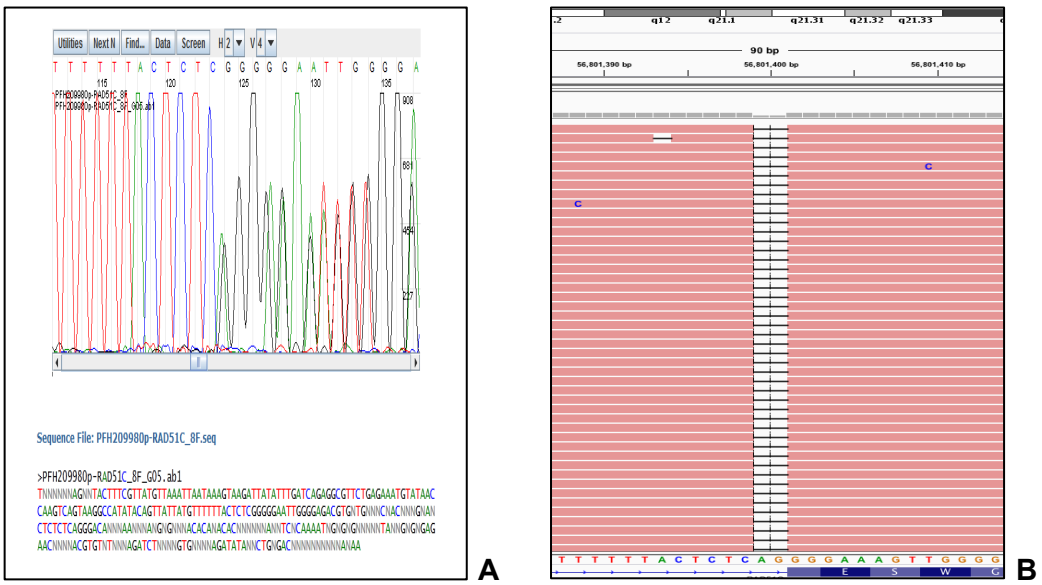


Figure 3.30 Sanger sequencing trace and NGS IGV generated image for variant RAD51C c.905-2delAG. *This figure shows images of the same RAD51C (905-2delAG) spice site mutation, A. Sanger sequencing trace and B. NGS IGV generated image. The NGS image clearly shows the 2 base pair deletion in the forward read direction. The grey boxes show the read depth at half the surrounding sequence showing that this is a heterozygous deletion.*

Figure 3.31 Sanger sequencing trace and NGS IGV generated image for variant *RAD51D* c.478C>T

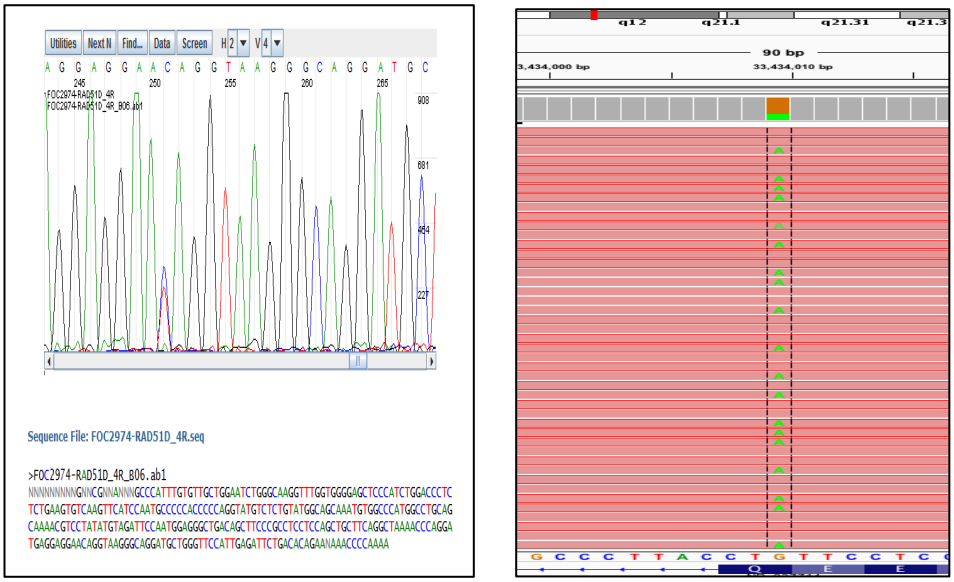


Figure 3.31 Sanger sequencing trace and NGS IGV generated image for variant RAD51D c.478C>T. *This figure shows images of the same RAD51D (c.478C>T) nonsense mutation, A. Sanger sequencing trace and B. NGS IGV generated image. The NGS image shows this is an heterozygous single base change with approximately half of the reads for each base (G = brown and A = green). The read is the forward direction and the gene is on the reverse strand.*

Figure 3.32 Sanger sequencing trace and NGS IGV generated image for variant *XRCC2* c.96delT

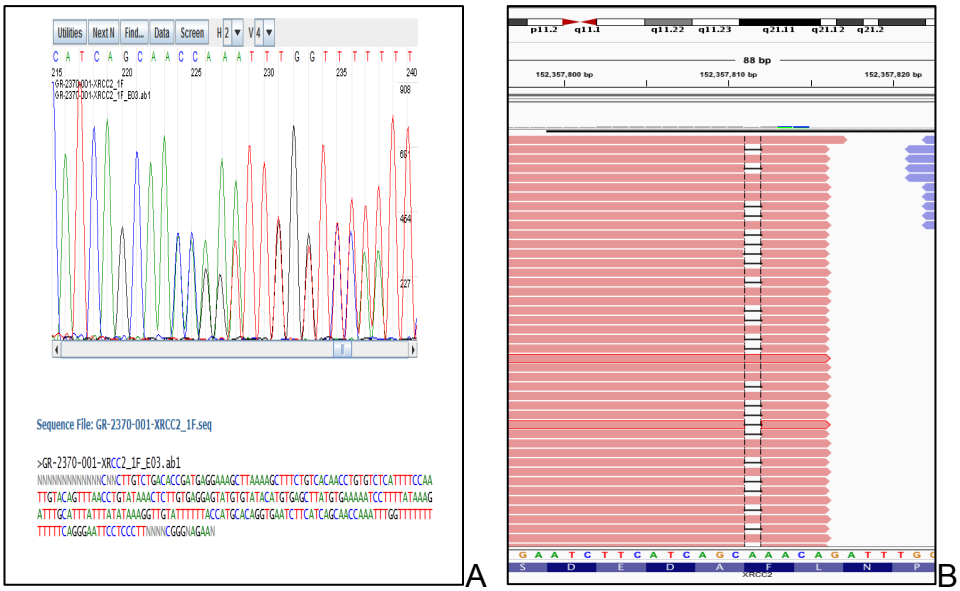


Figure 3.32 Sanger sequencing trace and NGS IGV generated image for variant XRCC2 c.96delT. *This figure shows images of the same XRCC2 (96delT) frameshift deletion, A. Sanger sequencing trace and B. NGS IGV generated image. The NGS image shows that approximately half of the forward reads contain the single base deletion and half do not suggesting a heterozygous single base deletion. Two reads show a red outline and this suggests they are poorer quality than the other reads.*

Figure 3.33 Sanger sequencing trace and NGS IGV generated image for variant *XRCC3* c.194-2A>G

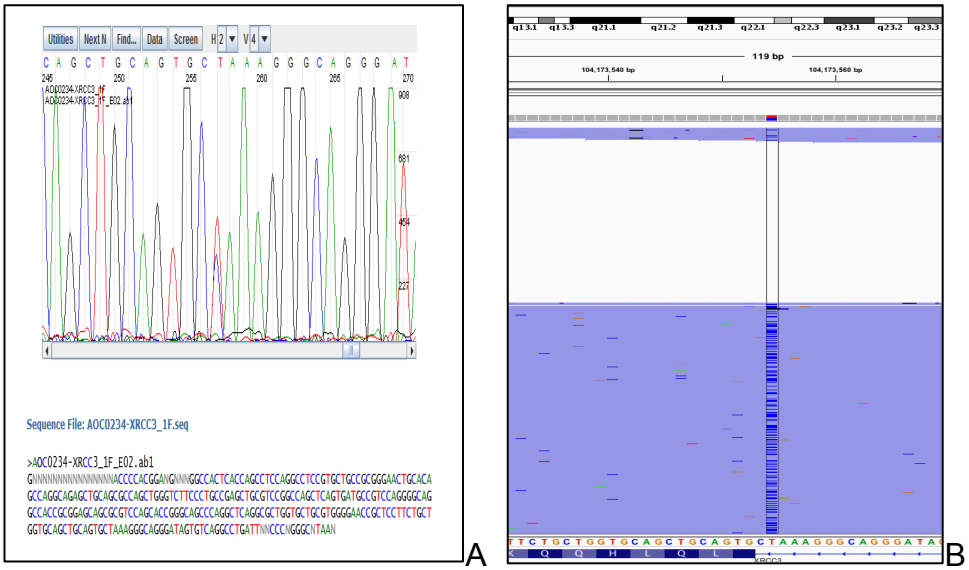


Figure 3.33 Sanger sequencing trace and NGS IGV generated image for variant XRCC3 c.194-2A>G. *This figure shows images of the same XRCC3 (194-2A>G) splice site mutation, A. Sanger sequencing trace and B. NGS IGV generated image. The NGS image has been set to 'squished' reads mode so that the reads containing the variant are visualised. There are clearly approximately half the number of reads with a T base (red) and half with the C base (blue). These reads are the reverse direction reads.*

3.18 Epidemiological data

Table 3.13 Epidemiological data for samples with predicted protein-truncating variants

Gene	Variant Type	Variant	Epidemiological Data							
			Race	Age at diagnosis	Histology	Grade	FIGO Staging	Final status at follow-up	Family history	Source of sample
RAD51B	Nonsense	c.489T>G	Unknown	62	Serous	Poorly differentiated	IIIC	Deceased 368 days	Unknown	AOCS
RAD51C	Frameshift deletion	c.497delT	White	52	Serous	Poorly differentiated	IIIC	Deceased 1716 days	No history of breast or ovarian cancer in family	MALOVA
RAD51C	Frameshift deletion	651_652del	White	64	Serous	Poorly differentiated	IIIC	Deceased 1031 days	No history of breast or ovarian cancer in family	AOCS
RAD51C	Nonsense	c.577C>T	Unknown	60	Other (i.e. not serous, endometrioid or mucinous)	Unknown	Unknown	Unknown	Maternal Aunt had ovarian and breast cancer	PFH
RAD51C	Nonsense	c.577C>T	Unknown	41	Serous	Moderate to poorly differentiated	Unknown	Unknown	Sister had ovarian cancer	PFH
RAD51C	Nonsense	c.955C>T	Unknown	40	Serous	Moderately differentiated	IIIC	Alive 2107 days	Father had breast cancer	AOCS
RAD51C	Nonsense	c.955C>T	White	74	Serous	Poorly differentiated	IIIC	Alive 2664 days	Sister had breast cancer	AOCS
RAD51C	Splicing	c.706-2A>G	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	One case of ovarian cancer (not FDR)	UKFOCR
RAD51C	Splicing	c.706-2A>G	White	64	Serous	Poorly differentiated	IIB	Alive 1418 days	Mother and sister had ovarian cancer	UKOPS
RAD51C	Splicing	c.905-2delAG	Unknown	52	Endometrioid	Unknown	IIB	Unknown	Mother ovarian cancer	PFH

RAD51C	Nonsense	c.312T>A	White	49	Serous	Poorly differentiated	IIIC	Alive 3062 days	No history of breast or ovarian cancer in family	AOCS
RAD51D	Frameshift deletion	c.565_568del	White	59	Serous	Moderately differentiated	IIIC	Deceased 2408	No history of breast or ovarian cancer in family	MALOVA
RAD51D	Frameshift deletion	c.565_568del	White	76	Serous	Poorly differentiated	IIA	Deceased 564	No history of breast or ovarian cancer in family	MALOVA
RAD51D	Nonsense	c.478C>T	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	One case of ovarian cancer (not FDR)	UKFOCR
RAD51D	Nonsense	c.620C>A	White	59	Serous	Moderately differentiated	IB	Alive 1246 days	No history of breast or ovarian cancer in family	UKOPS
RAD51D	Nonsense	c.898C>T	White	49 at interview	Control no ovarian cancer	NA	NA	NA	Sister had ovarian cancer	AOCS
XRCC2	Frameshift deletion	c.96delT	White	50	Serous	Unknown	Unknown	Deceased days unknown	One case of ovarian cancer (not FDR)	GRFOCR
XRCC2	Frameshift deletion	c.96delT	Unknown	40	Mucinous	Well differentiated	IA	Unknown	1st cousin had ovarian cancer (maternal side)	PFH
XRCC3	Splicing	c.194-2A>G	Unknown	64	Serous	Poorly differentiated	IIIC	Alive 3513 days	No history of breast or ovarian cancer in family	AOCS
SLX4	Frameshift deletion	c.2497_2498del	White	56	Serous	Moderately differentiated	IIIC	Deceased 1950 days	No history of breast or ovarian cancer in family	MALOVA
SLX4	Nonsense	c.4386_insAGGATGAACGAGGCCGC	White	61	Serous	Poorly differentiated	IIB	Alive 1153 days	No history of breast or ovarian cancer in family	UKOPS
SLX4	Nonsense	c.4386_insAGGATGAACGAGGCCGC	White	51	Serous	Moderately differentiated	IIIC	Deceased 2354 days	No history of breast or ovarian cancer in family	UKOPS

Table 3.13 Epidemiological data for samples with predicted protein-truncating variants. *FDR = First-degree relative. FIGO (International Federation of Gynaecology and Obstetrics) is the standard staging for epithelial ovarian cancer. The majority of cases are diagnosed at stage IIIC all of which are serous histology. The majority of cases are serous in histology, with only 1 mucinous and 1 endometrioid and 1 other. Age at diagnosis ranges from 40 to 76 years, with 11 cases being diagnosed under the age of 60 years and 8 over 60 years; the mean age at diagnosis is 57 years. Final status gives detail on whether the*

woman is alive or deceased at the final follow-up with the number of days from diagnosis to final follow-up. Survival ranges from 1 year to 9.6 years. Family history information is included giving details on which family member(s) are affected; 10 cases do not have family history of breast or ovarian cancer and a further 5 have family history that is not in a FDR; 5 cases have family history in at least 1 FDR.

Epidemiological data are given for samples with predicted protein-truncating variants. Table 3.13 gives epidemiological data for each of these samples including race, age at diagnosis, histology, grade, International Federation of Gynaecology and Obstetrics (FIGO) staging, which is the standard tumour staging system for epithelial ovarian cancer (Table 3.14 gives a description of FIGO staging for the tumour stages described in these data), final status at follow-up (deceased or alive) with number of days from diagnosis to final follow-up, family history and source of sample.

16 cases have known FIGO staging classification. In 10 cases diagnoses are made at FIGO stage IIIC (Table 3.14), with all of the IIIC staged tumours being of serous histology. 3 are made at stage IIB and 1 each at stage IA, IIA and IB. Stage IIIC is the more advanced stage with metastases evident beyond the pelvis. Only stages IA and IB are localised to 1 or both ovaries. 13 cases show tumour staging as extending beyond the ovaries at diagnosis.

For 19 cases histology is known; of these 16 are serous histology (84.2%), 1 is mucinous, 1 is endometrioid and 1 is classified as other (i.e. not described as serous, mucinous or endometrioid). This is in keeping with known prevalence statistics for serous histological subtype, which is estimated to be 80-85% in the Western world (Seidman et al 2004). In 16 cases tumour grade is known; of these 9 are the highest grade, poorly differentiated (56.2%), 1 is moderate to poorly differentiated and 5 are moderately differentiated. Just 1 case is the lower grade of well differentiated.

Age at diagnosis ranges from 40 to 76 years old with the mean age at diagnosis being 57 years. 11 cases are diagnosed in women under 60 years with 8 diagnoses over 60 years. Follow-up time ranges from 1 year to 9.6 years with 7 still living at final follow-up and 8 deceased at final follow-up. Of the women deceased, the shortest survival time was 1 year and the longest 6.6 years.

Family history data is available for the majority of cases. 10 cases have no family history of breast or ovarian cancer; 5 cases have family history of breast or ovarian cancer in relatives that are not first-degree relatives and 5 cases have family history of breast and/or ovarian cancer in 1 FDR or more. This means that 50% of cases with a predicted protein-truncating variant do not have family history of breast or ovarian

cancer; 25% have family history of ovarian and/or breast cancer in at least 1 FDR and 25% have family history of breast and/or ovarian cancer in a relative other than a FDR. For the 11 women diagnosed under the age of 60 years only 3 (27%) have family history of breast or ovarian cancer in a FDR; 6 (54%) have no family history and the remaining 2 (18%) have family history of breast and/or ovarian cancer in a relative other than a FDR. For the women with family history in FDRs diagnoses are made at 40, 41 and 52 years (mean age at diagnosis is 44.3 years). For the women without family history the youngest diagnosis is 49 and the oldest 59 (mean age at diagnosis is 54.3 years) and for the women with family history other than a FDR the two diagnoses were made at 40 and 50 years (mean age at diagnosis is 45 years).

Table 3.14 International Federation of Gynaecology and Obstetrics (FIGO)

Stage	Description
IA	Tumour growth is limited to one ovary, no ascites containing malignant cells. No tumour on external surface; capsule intact
IB	Tumour growth is limited to both ovaries, no ascites containing malignant cells. No tumour on external surface; capsule intact
IIA	Tumour extension and/or metastases to the uterus and/or fallopian tubes
IIB	Tumour extension to other pelvic tissues
IIIC	Metastasis evident beyond the pelvis >2cm in diameter and/or positive regional lymph node involvement

***Table 3.14 International Federation of Gynaecology and Obstetrics (FIGO)**. FIGO is the standard staging for tumour stages included in these data. Adapted from FIGO Committee on Gynecologic Oncology Int J Gynaecol Obstet 105 (1): 3-4, 2009.*

3.19 Ovarian cancer risks associated with predicted deleterious variants in candidate susceptibility genes

In order to evaluate the clinical significance of these findings it is necessary to evaluate the disease risks that are associated with deleterious mutations in these 6 genes for the population under study. These predicted deleterious variants are interpreted in the context of the available genetic epidemiological data and in the clinical characteristics of disease in the affected individuals. These data suggest that whilst family history of breast or ovarian cancer is not the overriding factor in ovarian cancer development it does appear to influence age of onset in that the mean age of diagnosis in women with family history is 10 years younger than those without family history. This information

has clinical relevance in that risk prediction and early detection strategies may be most beneficial for women with family history of breast and/or ovarian cancer beginning at the age of 40 (or less depending on the level of risk calculated within the family).

In the 5 youngest diagnoses (40 to 50 years), 3 of these have nonsense variants in *RAD51C*; interestingly all 3 diagnoses were made when women were in their 50th decade. 2 women have the same frameshift deletion in *XRCC2* of which 1 was 40 years and 1 was 50 years at diagnosis. These 2 cases with *XRCC2* variants are from the same small population from the Polish Family History study, suggesting either possible founder mutations or that these are distantly related to each other.

Of the 10 with FIGO tumour staging at the most advanced stage (IIIC), 1 has a predicted nonsense variant in *RAD51B*, 5 have predicted protein-truncating variants in *RAD51C* (2 frameshift deletions and 3 nonsense variants), 1 has a predicted splice site variant in *XRCC3* and 2 have predicted protein-truncating variants in *SLX4* (1 nonsense and 1 frameshift deletion). These data suggest that these particular gene variants may increase the risk of developing more aggressive disease; of the women in this study where FIGO staging is known, 62.5% are diagnosed at stage IIIC. All women diagnosed at stage IIIC have serous histological subtype tumours and the mean age of diagnosis in this subset is 57.1 years. Interestingly, of those women with IIIC stage tumours, 7 (70%) have no family history of breast or ovarian cancer, 2 have family history in a FDR (20%) and for 1 family history is unknown (10%).

3.20 Statistical analysis of data

3.20.1 Predicted protein-truncating variants

Odds ratios (OR) are calculated for predicted protein-truncating variant and predicted functional missense variants. OR is chosen over relative risk due to the low prevalence figures for these variants. OR is calculated to give an impression of the disease odds for those with a predicted deleterious variant, it measures the association between the presence of a positive variant and disease status. OR is calculated using the formula:

OR=(a/b)/(c/d) with 95% Confidence Intervals (CI) = OR ± 1.96 SE

A 2 X 2 contingency table is created as follows:

		Disease Status	
		Case	Control
Variant Status	Positive	a	b
	Negative	c	d

a = number of cases positive for a predicted deleterious variant
b = number of controls positive for a predicted deleterious variant
c = number of cases negative for a predicted deleterious variant
d = number of controls negative for a predicted deleterious variant

Upper and lower 95% confidence intervals are calculated from the standard error (SE) for the log OR as follows:

$$SE\ (logOR) = \sqrt{1/a + 1/b + 1/c + 1/d}$$

Thus, the 95% CI for the logOR is ±1.96 SE. To the find the 95% CI for the OR the antilog is found using exp(logOR ±1.96) SE.

To test the significance of the OR and ascertain the p-value, the z-statistic is used. This z-statistic is a measure of the standard deviation and from this the p=value obtained by looking up the z-statistic in standard tables.

Table 3.15 Calculated Odds Ratios for predicted protein-truncating variants

Predicted protein-truncating variants								
Gene	**Mutation status**	**Cases**	**Controls**	**Odds Ratio**	**95%CI (Lower)**	**95% (Upper)**	**Z statistic**	**P value**
RAD51B	Positive	1	0					
	Negative	1505	1130					
RAD51C	Positive	10	0					
	Negative	1496	1130					
RAD51D	Positive	4	1	3.0	0.33	26.94	0.984	0.325
	Negative	1502	1129					
XRCC2	Positive	2	0					
	Negative	1504	1130					
XRCC3	Positive	1	0					
	Negative	1505	1130					
SLX4	Positive	3	1	2.25	0.23	21.69	0.703	0.481
	Negative	1503	1129					

Table 3.15 Calculated odds ratios (OR) for predicted protein-truncating variants. *OR could only be calculated for RAD51D and SLX4 as these are the only genes where there is an occurrence of a predicted protein-truncating variant in the control group. 95% CI are large and neither of these are statistically significant at the p=0.05 level.*

Predicted protein-truncating variants in all 6 genes are detected in cases only for *RAD51B*, *RAD51C*, *XRCC2* and *XRCC3*. In *RAD51D* and *SLX4* each has one predicted protein-truncating variant in a control sample (Table 3.15). OR can only be calculated for *RAD51D* and *SLX4* because there are no predicted protein-truncating variants in controls in the other 4 genes. Neither of these is statistically significant and the 95% CI are large due to the rare prevalence of these variants and the sample size in this study. Increasing statistical power would require larger sample sizes. This suggests that the variants in *RAD51B*, *RAD51C*, *RAD51D*, *XRCC2*, *XRCC3* and *SLX4* are very rare. The likely risk curves for variants in these genes compared to *BRCA1* and *BRCA2* are represented in Figure 3.34.

3.21 Discussion

In this study a high throughput targeted next generation sequencing (NGS) approach is used to evaluate the prevalence and penetrance of germline functional mutations of a series of candidate susceptibility genes in 1506 ovarian cancer cases compared to 1130 controls. The NGS technology used here enables study sample sizes to reach levels required for the detection of rare or very rare variants in population-based case control studies. The Fluidigm Access Array system is used as a high-throughput, rapid and affordable method for NGS target enrichment and library preparation. This effectively removes the library preparation step previously required for sequencing on the Illumina HiSeq2000

3.22 Evaluation of the high-throughput NGS approach established in this study

3.22.1 Target enrichment and library preparation

The simple and streamlined workflow of the Fluidigm Access Array platform makes it ideal for use in large-scale research projects as well as in clinical diagnostic laboratories. The Long Range PCR (LR-PCR) method (previously described in Chapter 2) involves several weeks of LR-PCR experiments as target enrichment to amplify 1 gene in 11 samples. The standard Illumina library preparation kit (in 2010) requires several days for completion for all 11 samples that are run in 1 Illumina flow cell lane. In this study 48 samples are prepared in 1 day, including both target enrichment and library preparation for the coding region of all six genes. Therefore, in just 8 days all 384 samples that are run in 1 Illumina flow cell lane are prepared. This platform produces sequence-ready libraries with Illumina sequencing adapters and primers attached. The multiplexing barcode index sequences are also included in this one-day protocol. Not only does this result in rapid target amplification and library preparation, it also means huge reductions in costs, estimated at around £4 per sample for all regions sequenced, then adding a cost for analysis results gives a final price of around £15 per sample. This cost is likely to continue to reduce as multiplexing levels are increased further. However, it is important to note that for the diagnostic setting, whilst this is a good streamlined method for target enrichment and library preparation, in order to reach the lower costs there is a requirement for preparing large numbers of samples, which a diagnostic clinic may not receive on an on-going basis.

This study employs the use of a PCR based targeted capture approach to select the coding region only for the 6 target genes of interest. One limitation in this approach is that this method does not allow for the detection of larger genomic rearrangements (i.e.

large deletions and duplications). This is because this PCR based targeted capture method creates overlapping amplicons of around 200 bp each. Therefore, if a large deleterious deletion is present, for example of a few kilo base pairs (Kb), the PCR primers will not be annealing to the mutant copy of the genes, unless they span a breakpoint that lies within the coding region. This problem can be circumvented using capture techniques that do not rely on short PCR fragments. In a paper by Walsh et al (2010) they assess the efficacy of genomic capture as a target enrichment method to enable the isolation of 21 genes they expect to be implicated in inherited ovarian cancer. Their method includes the coding sequence as well as the intronic sequences plus an additional region of 10 Kb upstream (5') and downstream (3') for each gene. This way they are able to detect 4 large rearrangements in *BRCA1* and 2 in *BRCA2*. Of these, they describe 5 as deletions and 1 as a duplication. They achieve this by calculating the read depth at each base pair for individual samples and make a comparison with the mean read depth for the same bases for all samples. If the calculation is less than 60% or over 140% of the mean depth for all samples in the study then they assume that these samples are not diploid. If less than 60% then they suggest there is a large deletion and where mean depth is over 140% they suggest a duplication. Walsh et al (2010) use Sanger sequencing to validate the breakpoints of large deletions and duplications and conclude that 12.5% of *BRCA1* mutations are large deletions that result in inactivation of BRCA1 protein. However, the main argument for not using a genomic capture method in this study is due to the high costs for library preparation and the large quantity of DNA required; Walsh et al (2010) report that 3 µg of genomic DNA are required for each individual's library preparation. This compares to the 100ng required per sample using the protocol in this study.

3.22.2 Sequencing quality controls (QC) – sequence coverage

An important factor affecting sequencing data quality is the sequence depth of coverage or read depth; that is the number of times each individual base is sequenced. The requirement for varying depths of coverage specifically relates to the allele frequency of variants under investigation. Very common variants, i.e. those with minor allele frequencies between 5% and 50% (Cirulli & Goldstein 2010) are relatively easy to detect (see Figure 3.34 in Results) especially with large samples sizes in this study. Therefore, relatively low coverage (e.g. 10X) will be sufficient to detect the vast majority of these variants. However, highly penetrant variants such as those in *BRCA1* and *BRCA2* are very rare in the general population. In a paper by Whitmore et al (2004) they estimate minor allele frequencies of 0.24% in a white European non-Ashkenazi Jewish population. The abundant different non-founder functional variants identified in

these genes, means that the frequency of each individual variant in the population is extremely rare. Sequence coverage >50X is often regarded as the lower threshold in sequencing data to be sure of detecting the alterations of this frequency. For recently published population based studies that have identified susceptibility genes for ovarian cancers (e.g. *RAD51C* and *RAD51D*), the frequency of mutations in these genes appears to be even lower than for *BRCA1* and *BRCA2* (<1% in familial ovarian/breast cancer cases) and so depth of sequencing coverage becomes even more of an issue.

The Illumina HiSeq2000 NGS approach sequences PCR amplified DNA molecules and this can introduce amplification bias in sequencing data resulting in false negatives. Thus, the appropriate level of coverage is required to enable the accurate detection of variants. The level required depends both on allele frequencies and read filtering parameters. In terms of detecting heterozygous variants (i.e. where there will be 50% of reads for each allele) a read depth of 1X will mean that only 0.5 of variants will be detected. Moreover, as read depth increases the probability of detecting variants also increases; as reads are filtered out though due to quality scores, then even a read depth of 15X will miss a proportion of variants. In this study a read depth of 30X is considered the minimum level for confident variant detection, however in the diagnostic clinic this level is considered to be 50X to reduce further the possibility of a variant call error. Morgan et al (2010) look at read depth and calculate that this positively correlates with variant detection. In their LR-PCR study they simulate lower read depths and calculate that at a minimum read depth of 50X they are confident that no variants are missed.

For the current study, the overall coverage across samples is shown to be relatively even both between samples and throughout the regions sequenced. Compared to LR-PCR, this is a far superior approach since in the LR-PCR method some regions have very low or no coverage and the minimum threshold for mutation detection (i.e. 30X) is not always attained. In this study a mean sensitivity of 94% >30X depth is observed across all 6 genes; and the overall median coverage for samples is 2,264X. This suggests that a larger region could be included to still reach acceptable read depth for accurate mutation detection. A small proportion (1.93%) of samples failed (n=51 with read depths <30X for 80% of the sample); this may be due to DNA quality and as such, a level of failure can be anticipated. The de-multiplexing step is proven to work well as verified through positive control samples that are accurately identified.

3.22.3 Primer chopping

The sequencing primers are chopped out during the analysis meaning that bases are removed from one end of reads. Following this the read co-ordinates are adjusted, but this does not require re-alignment of reads. This can pose a problem if bases are not incorporated, during sequencing reactions, at the end of primers or if additional bases are added. If there was a deletion or insertion in a region to which a primer annealed then this could result in an alteration of the position of reads following primer chopping.

3.22.4 Variant detection sensitivity

Sanger sequencing is still considered the 'gold standard' in variant detection (Davidson et al 2012). However, this method is limited in terms of cost and throughput, and is not suitable for the large-scale epidemiological based approaches that include the thousands of subjects that are needed to detect rare susceptibility variants in the population. It is for this reason that NGS is an attractive alternative and therefore, an evaluation in the variant detection sensitivity and specificity of the approach is essential.

Many variants are removed at the read filtering stage. This is due to minimum thresholds at minor allele frequency rates that allow for confident variant detection. At a read depth of >15X the minor allele frequency must be a minimum of 40% to remain in the data set; at a read depth of >30X the minor allele frequency must be a minimum of 30% to remain in the data set. Variants not meeting these criteria are filtered out. The performance of lanes, in terms of the proportion of remaining variants post-filtering ranges from 34.4% (lane 4) to 97.7% (lane 2). In lanes 1, 3, 5, 6 and 7 there are 94.5%, 89.8%, 87.9% and 92% respectively remaining variants post-filtering. Looking at performance of samples in terms of read depth, between 93% and 95% of samples are sequenced at >30X depth across all 6 genes, suggesting very good rates of sensitivity. That is, there is a very high probability that the sequencing will detect all variants.

The mutation detection sensitivity can be viewed as an estimate of the false negative rate i.e. of the number of mutations missed by the sequencing approach. This is difficult to calculate since not all samples can be analysed by Sanger sequencing. However, an evaluation can be made using positive control samples. These positive controls are 'spiked' into the experiment and blinded for the purpose of calculating sensitivity as well as for assessing the multiplexing accuracy in the study as a whole. Variant detection

sensitivity is defined as the proportion of known mutation positive controls correctly identified as positive by NGS. The following formula calculates the sensitivity rate:

$$\frac{\text{No. True positive variant calls}}{\text{No. Positive variant calls + false negatives}}$$

This gives a measurement of the true positive rate. In this study there are 6 positive controls and 5 of these are accurately detected giving a mutation detection sensitivity rate of 83.3%. The variant not detected is in a region of zero depth coverage. Therefore, where there is adequate coverage mutation detection sensitivity is 100%, however, the coverage for all genes >30X was 94% so this is the maximum possible sensitivity for the whole study. The results section refers to these coverage data. In the clinical setting false negatives are important to guard against, as we will not Sanger sequence all samples in a study or clinical setting. We need to strive for 100% sensitivity in both research and diagnostic setting.

In addition, it is possible to estimate the false negative rate in a study by making comparisons with known variants from a public database such as dbSNP. The false negative rate can be calculated as the proportion of known variants that should be detected these data, but are not. Illumina suggest that this can also be calculated as true positive rate subtracted from 100. Harismendy et al (2009) compare Illumina sequencing data to Sanger sequencing data to calculate false negative rates and in their data they find zero false negative variant calls. In addition, they find 100% agreement between SNP genotyping data and Illumina NGS data. Furthermore, they suggest that at simulated low coverage, the rate of false negatives is still minimal.

3.22.5 Mutation detection specificity

Mutation detection specificity is defined as the true negative rate. This is not possible to calculate as we cannot Sanger sequence all genes in all samples. Therefore, it may be most appropriate to assess the false positive rate. This is calculated using Sanger confirmation of a proportion of the variants detected. If variant detection specificity is viewed as the rate of true negative variant calls then the following formula would give this figure:

$$\frac{\text{No. True negative variant calls}}{\text{No. True negative variant calls + false positive variant calls}}$$

The study includes a number of non-template controls, however this does not give an accurate figure for the specificity as the non-template controls only test for contamination during PCR or sequencing. In this study the large numbers of samples mean that it is prohibitive in terms of cost and time to validate all samples using Sanger sequencing. There are a number of studies that have examined the false positive rates for variant detection in NGS studies. Ozcelik et al (2012) use LR-PCR and NGS to sequence the whole of *BRCA1* and *BRCA2* validating all 12 patient DNA samples with Sanger sequencing. They report that both false negative and false positives are zero in their study.

In this study, 28 NGS variants are Sanger sequenced for confirmation and 24 of these are detected with the remaining 4 being false positives. This means that 85.7% are confirmed. In the clinical setting false positives are less worrying than false negatives as currently all NGS positives are confirmed using Sanger sequencing, therefore, false positives would be eliminated.

3.23 Evaluation of study design

3.23.1 Targeted candidate gene approach versus whole exome sequencing

Other studies suggest that candidate gene approaches are not finding rare variants in moderate penetrance genes (Thompson et al 2012) and that whole exome sequencing should instead be used to identify causal rare variants. Whilst there have been substantial increases in capacity and throughput of NGS technologies in recent years there are still not enough high-throughput whole exome sequencing studies to support this approach; particularly of very large case-control studies with the aim of detecting rare or very rare variants. The place for exome sequencing may be in family-based studies, which can be informative of candidate genes for follow on case-control studies. The targeted candidate gene approach is still likely to be the most appropriate method for detecting rare or very rare variants in ovarian cancer susceptibility genes. It may be appropriate in breast cancer to perform full exome sequencing on small pedigrees of breast cancer families and then re-sequence the genes discovered in large case-control studies. Since we can draw on our existing knowledge of the biology of ovarian cancer it is possible to choose panels of likely gene candidates. In order to increase the probability of finding rare variants in novel ovarian cancer susceptibility genes the best approach is likely to involve the most appropriate sample set, i.e. a genetically enriched sample set of women at very high risk of developing ovarian cancer. The increasing capacity of NGS technology and improvements in bioinformatics and data analysis is allowing for the increase in size of these candidate gene panels. Within 12

months it is possible to double the number of genes under investigation from 6 to 12 genes.

An advantage of targeted re-sequencing compared to whole exome or whole genome is in the number of samples that can be run in one experiment resulting in more cost effective experiments and furthermore, the depth of coverage is far higher in a targeted candidate gene approach. Exome sequencing results in much lower depth of coverage and this is very likely to result in missing variants that are relevant (Walsh et al 2011). To characterise novel cancer susceptibility genes, exome sequencing could be most useful, as this would allow for narrowing down the number of genes interrogated. However, to ensure that variants are not missed or for use in a clinical setting, a targeted re-sequencing approach is likely to be the most appropriate, since this will increase sensitivity as well as allow for increasing sample throughput to levels where resulting data can be confidently applied to a wider population. In addition, in the genetic clinic targeted re-sequencing of a panel of ovarian cancer genes can be screened with the aim of offering an 'ovarian cancer risk' percentage.

Exome sequencing has many caveats compared to targeted re-sequencing studies. Whilst there is background noise in both exome and targeted sequencing studies, whole exome sequencing will have much higher levels of background noise simply because there are many more genes under investigation and many people have variants in many genes that appear not to be harmful. This is particularly the case with complex diseases such as cancer. Additionally, when searching for genes in autosomal dominant diseases then a higher coverage is required to confidently identify mutations. The reason for this is that the search will be for heterozygous rather than homozygous variants and thus the need for double the level of coverage. It can be estimated that to detect ~95% of single nucleotide variants in an exome sequencing study at least 20X depth is required; below this level of depth, studies are likely to miss many of the causative variants.

3.23.2 Advantages and disadvantages of population based case control studies

One of the main confounding issues in population-based studies is that of population stratification, in which allele frequencies differ between sub-populations. This is due to the diversity in ancestry between individuals where minimal mating has occurred between these sub-populations meaning that allele differences are not shared between these groups (Li et al 2010, Foulkes 2009). This study includes data on race and the samples are drawn from ovarian cancer registries and other ovarian cancer studies

worldwide. To fully define the risk associated with gene variants in these types of studies, it would be necessary to correct for this type of confounding to guard against false-positive effects due to population stratification. Whilst bioinformatics correction methods are available for this purpose, another way to surmount this caveat is to perform replication studies in which cases and controls are selected randomly from populations. Since, data from many large international case control studies is likely to be combined to allow for full determination of ovarian cancer risks in these genes, population stratification should not result in false-positive effects.

3.24 Genetic variant prevalence and characteristics

In this study the combined prevalence of a deleterious mutations in all 6 genes is 1.33% in ovarian cancer cases compared to 0.09% in the healthy controls. For the purpose of this analysis, deleterious mutations are classified as frameshift, nonsense and splice site alterations predicted to result in protein truncation. Missense mutations are identified in the 6 genes, although their functional significance is unclear, and variant prediction programs such as SIFT and Polyphen-2 are limited in their ability to predict function with confidence. However, when these variants are combined, the prevalence of predicted functional missense variants across all six genes in the set is 42.63% (cases) and 34.5% (controls). This suggests that most of these predicted missense variants are likely not to result in non-functional protein. Certainly, those missense variants that occur at high frequency in cases and controls can be eliminated as non-pathogenic. In those missense variants that are low frequency in controls it could be relevant to perform functional assays to assess the impact of these variants on the protein product and then calculate the combined prevalence in cases and controls.

In 2011, The Cancer Genome Atlas (TCGA) published data on mutational analyses of 316 high-grade serous ovarian carcinomas (HGSOC). They report that half of these tumours are functionally deficient in one of the homologous recombination (HR) genes either by germline or somatic mutation or by hypermethlyation leading to gene silencing. In terms of somatic mutations >3% of HGSOC tumours have mutations in *BRCA1* or *BRCA2*, with a further 6 genes identified as significantly mutated in tumours. 2 of these can be ruled out (FAT tumour suppressor homolog (*FAT3*) and Gamma-Aminobutyric Acid (GABA) A receptor Alpha 6 (*GABRA6*) as they do not appear to be expressed either HGSOC or fallopian tube normal tissue. The 4 other genes include *RB1* (a known tumour suppressor gene), neurofibromin 1 (*NF1*), cyclin-dependent kinase 12 (*CDK12*) and CUB and Sushi multiple domains 3 (*CSMD3*). Of these, 2

(*NF1* and *CSMD3*) do not appear to be expressed in ovarian or fallopian tube tissues (Online Mendelian Inheritance in Man - OMIM), leaving *CDK12*, which TCGA report has 5/9 predicted protein-truncating mutations with the remaining 4 predicted missense variants. In addition, TCGA analyse somatic mutations detected in their data with those in OMIM and in the Catalogue of Somatic Mutations in Cancer databases. From these analyses they suggest that other rare variants may be key players in ovarian, including *BRAF*, *PIK3CA*, *KRAS* and *NRAS*, interestingly these are all oncogenes, which are uncommon in inherited cancer syndromes (discussed in Chapter 1). Certainly, for future sequencing studies, candidate genes can be centred on the homologous recombination related genes, as these appear to be highly relevant in ovarian cancer.

The majority of the predicted protein-truncating variants detected in this study are novel, with the exception of 3 variants that are previously reported in literature as associated with ovarian and/or breast cancer. These are all in *RAD51C* and include a nonsense variant (c.577C>T) and a splice site variant (c.905-2delAG); both are published by Coulet et al (2012) and another splice site (c.706-2A>G) is published by Loveday et al (2012).

Of the predicted protein-truncating variants in these data 3 variants (2 nonsense and 1 splicing) are identified in more than one case. For each, the cases are apparently unrelated, but from the same geographical region. Variant *RAD51C* c.577C>T is a nonsense mutation in exon 4 and the two cases in which this is detected are both from Poland with family history of ovarian cancer. There are only 96 polish family history cases in the whole study meaning that the prevalence in the study for this variant is 2.1%; suggesting that this may be a novel founder mutation in this population or that these two cases may be distantly related to each other. To confirm a founder mutation a series of cases from the same population would need to be analysed for this variant. Variant *RAD51C* c.955C>T is a nonsense mutation in exon 7 and both of these cases are detected in Australian ovarian cancer cases. One of the splice site alterations (*RAD51C* c.706-2A>G) is detected in 2 UK cases from different studies one has family history of ovarian cancer in a FDR and the other has family history in a relative other than a FDR.

3.25 Analysis of clinical relevance of study findings

In cases with predicted protein-truncating variants and staging information (n=16) 62.5% are diagnosed at the more advanced stage of IIIC where distant metastases are already evident; and 81.2% are diagnosed with staging beyond the pelvis, suggesting that these variants may be relevant in more advanced disease. In cases with known histology (n=19) 84.2% are serous histology, which further suggests that homologous genes are highly relevant in HGSOC. Future studies may concentrate on both homologous recombination genes and HGSOC, especially since homologous recombination defective tumours are identified as sensitive to PARP inhibition (Mukhopadhyay et al 2012).

An earlier age of onset appears to be suggestive of the presence of a gene mutation; of those cases with predicted protein-truncating variants the age at diagnosis ranges from 40 to 76, with mean age at diagnosis being 57 years. This may be relevant clinically in the management of women with ovarian cancer in that for women diagnosed under 60 years it may be relevant to undertake genetic screening analysis as these tumours may have a sensitivity to PARP inhibition.

Interestingly, 50% of cases with predicted protein-truncating variants have no family history of breast or ovarian cancer, 25% have family history in a relative other than a FDR and 25% have family history in a FDR. This suggests that factors other than family history are involved in these cases. However, of particular importance is the observation that the mean age at diagnosis for those with family history is 10 years younger than those without. This factor is relevant in clinical management suggesting risk prediction and early detection measures should begin earlier for women with family history of breast and ovarian cancer. Indeed, depending on the level of risk calculated within the family, early screening or genetic testing could be recommended before the age of 40 years.

In a paper by Alsop et al (2012) they calculate the frequency of BRCA mutations in a cohort of women with non-mucinous ovarian cancer and analyse their response to treatment. They report that 44% of women with a *BRCA1* or *BRCA2* mutation do not have family history and since these women appear to have specific treatment response and survival, Alsop et al (2012) suggest all women presenting with non-mucinous ovarian cancer should be offered genetic testing for *BRCA1* or *BRCA2* routinely even without family history.

To fully define the risk of ovarian cancer caused by germline mutations in these genes would require extremely large numbers of cases and controls and this is due to the low frequency of these variants within the population. Loveday et al (2011) suggest this frequency to be around 0.1% in the population. The figure 3.34 is a graphical representation of the association between allele frequency and risk of ovarian cancer, which suggests that the more rare a variant the larger the effect of that variant. This figure can also give an indication as to the likely method of detection for finding these variants, in that common low-risk variants would most efficiently be detected via large Genome Wide Association Studies (GWAS), rare moderate-risk alleles are likely to detected via large studies of cases and controls in re-sequencing studies of defined candidate genes in specific disease phenotypes (e.g. HGSOC), perhaps alongside enrichment for family history and age of onset. Although, caution must be noted in calculation of odds ratios and inferring disease risk, especially in a single case control study, the odds ratio is given as an impression of disease association only. In this study odds ratios are calculated for predicted-protein truncating variants only as the predicted missense variants show high frequencies in controls in addition to cases, suggesting that they are not pathogenic. The low frequency of variants means that neither *RAD51D* nor *SLX4* odds ratios are statistically significant. Figure 3.34 places *RAD51C* amongst the rare moderate-risk alleles and if *RAD51D*, *XRCC2* and *XRCC3* were also placed on this diagram they would inhabit a similar region on the graph.

Figure 3.34 Minor Allele Frequencies versus relative risk

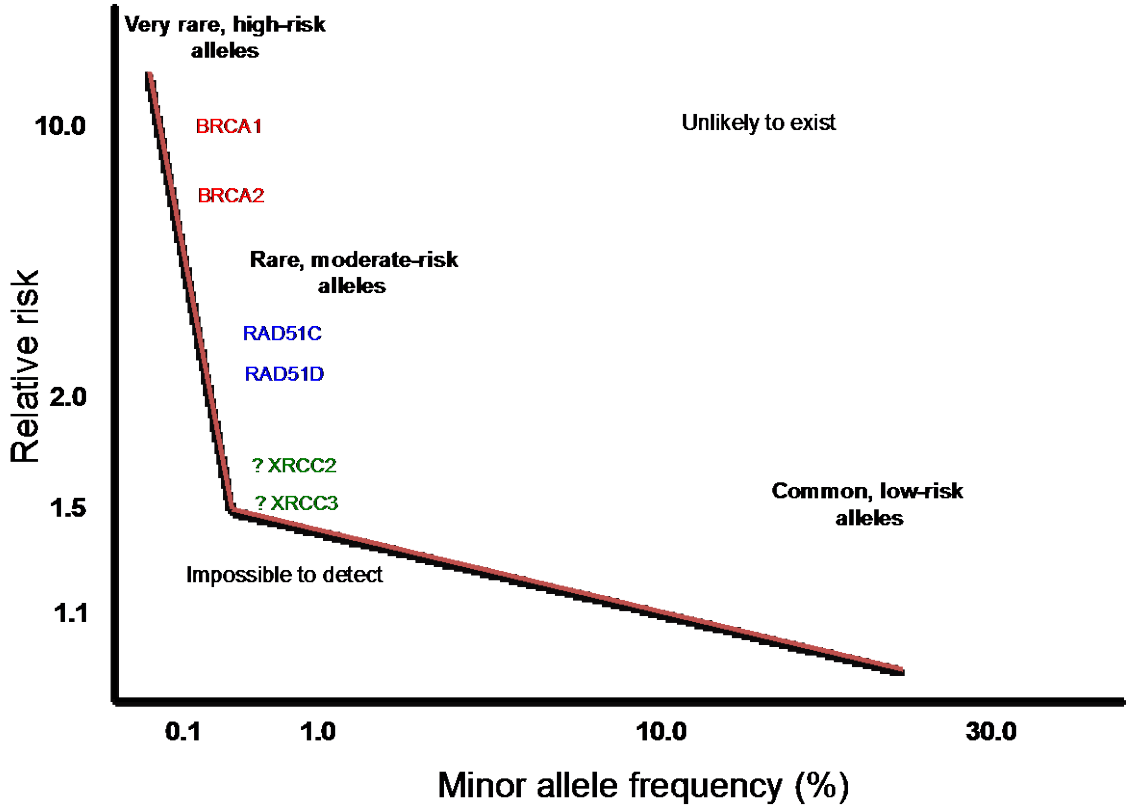


Figure 3.34 Minor allele frequencies versus relative risk. *This image is re-drawn and adapted from Harris & McCormick Nat. Rev. Clin. Oncol. 7, 251–265 (2010) with RAD51C and RAD51D added as new ovarian cancer rare moderate-risk susceptibility alleles. If XRCC2 and XRCC3 were new susceptibility genes then they would be situated close to RAD51D in terms of minor allele frequency, but below them in relative risk. The very rare low-risk alleles are impossible to detect, as the number of cases required would be too large. It is very unlikely that more common high-risk alleles exist, as these would have already been discovered. It is likely that many more rare, moderate-risk alleles exist and these are best discovered using high throughput re-sequencing approaches in large studies of cancer cases and controls.*

3.26 Conclusion

In conclusion, returning to the research questions outlined in the 'Introduction', the Fluidigm Access Array system, does offer a viable solution to the bottleneck seen in the library preparation step in Illumina sequencing. This novel approach is an accurate, rapid and cost effective method for the accurate detection of gene variants that predispose to an increase in ovarian cancer risk. This is demonstrated in the speed with which libraries are prepared, the depth and evenness of coverage, and the mutation detection sensitivity and specificity. This system accurately characterises the frequency of mutations occurring in the study population in the six genes under investigation in this study. To this end the four aims of this research are successfully addressed. This confirms the significance of *RAD51C* and *RAD51D* in ovarian cancer.

Whilst, no novel cancer susceptibility genes are among these candidates, this high-throughput system is now fully established and evaluated to allow for follow on studies that will have the capacity to identify novel cancer susceptibility genes. This research also suggests that these approaches could soon be introduced into the clinical setting meaning that a wider population of women could be offered genetic testing for early detection and prediction of risk. The epidemiological and histological data presented here in combination with detected gene variants suggests that specific risk factors may indicate that a genetic variant is involved and that testing for validated gene variants may be a routine part of clinical management of cases. This can be put into place once these gene variants are fully defined in terms of risk of ovarian cancer.

Chapter Four

A characterisation of 9 ovarian cancer susceptibility genes in unaffected women from high-risk breast-ovarian cancer families

4.0 Introduction

The connection between DNA damage and tumourigenesis has been intricately studied and well documented (Bernstein et al 2002). When DNA damage sensors detect DNA damage, the cell has a choice of going into either apoptosis or a temporary pause of the cell cycle during which time the damage is repaired. The biological mechanism that switches a cell between a pause in the cell cycle or apoptosis is crucial in cell protection from uncontrolled replication of damaged DNA and subsequent development of cancer. During the pause in the cell cycle where an attempt is made to repair damaged DNA the cell either continues with the damage repaired or goes into apoptosis, if the damage is too great to allow for its repair. If these cells are impervious to apoptosis this results in an increased production of cells with damaged DNA and a state of global genomic instability ensues; this leads to an elevated chance of carcinogenesis.

This study uses the same NGS analysis as the case-control study to analyse a panel of candidate genes. These genes have roles in DNA repair via homologous recombination; are downstream targets of key players in the homologous recombination pathway (for example *BRCA1* or *BRCA2* related genes) or are part of the Fanconi anaemia-BRCA DNA repair pathway.

4.1 DNA damage and the Fanconi anaemia pathway

The Fanconi anaemia pathway is crucial in the repair of DNA interstrand crosslinks. Fanconi anaemia is an autosomal recessive condition that results in congenital defects and malfunction of bone marrow as well as an increased susceptibility to cancers. This genetic disorder is very rare with an occurrence of around 1 in 100,000 births (D'Andrea & Grompe 2003). Some communities appear to show a higher incidence of the syndrome (for example, Ashkenazi Jewish and Afrikaners from South Africa). The study of this syndrome has increased understanding of the genetic susceptibility to cancers amongst those that do not have the genetic disorder, namely those related to

inherited mutations in *BRCA1* or *BRCA2*; as these genes are intimately related and involved in the response to DNA damage. Most interestingly, the Fanconi anaemia pathway also includes genes, the protein products of which are important in the activation of DNA damage checkpoints and temporary cell cycle cessation whilst repair takes effect. These include DNA damage sensor *ATM* in addition to *NBS1* and *RAD51*. The biological mechanism of these interrelated proteins sheds relevant insight on and direction towards appropriate candidate genes as targets for investigation in this high-throughput study.

4.1.2 The intricate relationship between tumour suppression and the Fanconi anaemia pathway

Fanconi anaemia patients with monoallelic mutations in the Fanconi anaemia pathway genes have an extremely elevated predisposition to many cancers including acute myeloid leukaemia (AML) and squamous cell carcinoma. Four downstream players in the Fanconi anaemia pathway function in DNA repair via homologous recombination and result in an increased risk of development of ovarian and/or breast cancer. *BRCA2* is also known as *FANCD1*. Interestingly, mutations in *RAD51C* exhibit a Fanconi-like condition and this may soon be referred to as *FANCO* (Kottemann & Smogorzewska 2013). Essentially, the predominant relevance of Fanconi anaemia pathway genes is in the interaction of these genes with *BRCA2*, meaning that homozygous mutations in Fanconi anaemia genes appear to affect the action of *BRCA2* in a way to render *BRCA2* inactive.

4.2 Rationale for choice of genes in this study

The 4 known genes in this study (*BRCA1*, *BRCA2*, *RAD51C* and *RAD51D*) are already identified as known ovarian and/or breast cancer susceptibility genes. The 5 candidate genes are chosen as they are either related to these 4 genes or the DNA double strand-break repair mechanism via homologous recombination. These 5 genes are *RAD51B*, *PALB2*, *NBN*, *BRIP1* and *BARD1*. The biology, structure and function of *BRCA1*, *BRCA2*, *RAD51B*, *RAD51C* and *RAD51D* are previously described in chapters 2 and 3 of this thesis. The additional 4 candidate genes are described here.

4.2.1 The structure and function of *PALB2*

The gene partner and localiser of *BRCA2* (*PALB2*) NCBI Genbank accession number NM_024675.3 (OMIM 610355) is also referred to as *FANCN* gene and forms part of the Fanconi Anaemia (FA) pathway. It is located on chromosome 16 at 16p12.2 with genomic co-ordinates:chr16:23614483-23652678. The proteins PALB2 and BRCA2 co-localise forming nuclear foci, with PALB2 assisting in the stabilisation and localisation of BRCA2 within the nucleus and facilitating BRCA2 in its roles of checkpoint control and repair by homologous recombination. Tishcowitz & Xia (2010) report that harbouring heterozygous deleterious changes in this gene results in an increased risk of both breast and pancreatic cancer. *PALB2* is involved in tumour suppression as part of the BRCA complex of DNA repair via homologous recombination. The protein product of this gene also appears to bind BRCA1 in addition to BRCA2. The diagram (Figure 4.1) below describes the structure of *PALB2* and which regions are likely to bind *BRCA1*.

The gene MORF-related gene 15 protein (*MRG15*) is also referred to as mortality factor 4 like 1 and is part of the histone acetyltransferase complex (HAT); the encoded protein functions in the transcription activation of specific genes, the relevance here is that it has a role in the localisation of PALB2/BRCA2/RAD51 to DNA damage foci.

Figure 4.1 The structure of *PALB2* with binding regions for interacting genes

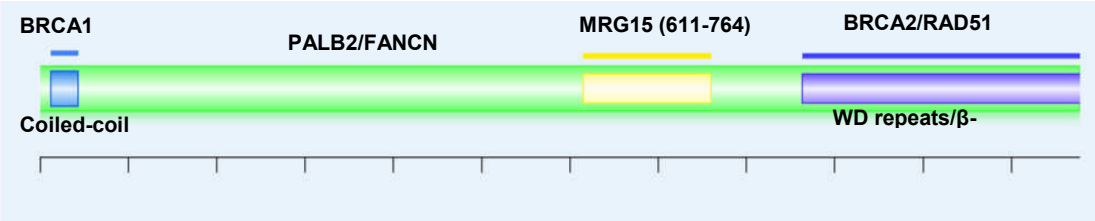


***Figure 4.1 The structure of PALB2 with binding regions for interacting genes**. This diagram demonstrates binding regions and relationships between PALB2 and BRCA1, BRCA2, RAD51 and MRG15. Adapted from Tischkowitz & Xia (2010) Cancer Res; 70(19) October 1, 2010*

Tischkowitz & Xia (2010) report on the model in which the BRCA DNA repair complex assembles at regions of DNA double-strand breaks. This core complex is composed of BRCA1/BRCA2/PALB2 as well as BRCA1 and BRCA1 C-terminus (BRCT) binding domain partners. These protein partners are likely to include BRIP1, FANCJ, coiled-coil domain-containing protein 98 (CCDC98-RaP80) or retinoblastoma binding protein 8 (RBBP8). There are two probable biological mechanisms by which BRCA1 is employed at detected regions of DNA damage. The first possibility is that BRCA1 interacts with the MRN complex (which includes MRE11/RAD50/NBS1) and the

second is that BRCA1 binds the CCDC98/RAP80 complex. One or other of these is the route via which PALB2/BRCA2/RAD51 associates with BRCA1 at DNA damage sites.

Figure 4.2 Schematic representation of *PALB2* (Partner and localiser of BRCA2) gene

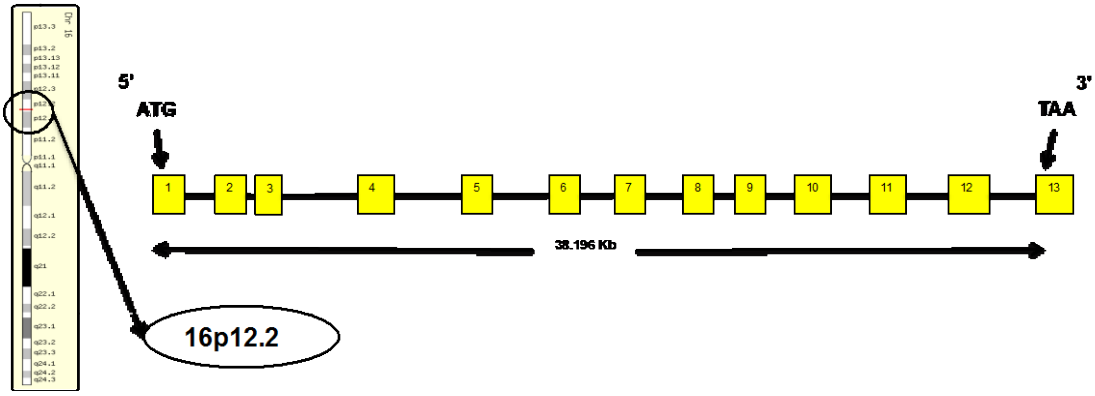


***Figure 4.2 A schematic representation of PALB2**. PALB2 has 13 exons all of which are coding exons and includes a genomic region of 38,196 bp.*

Deleterious mutations in *PALB2* are associated with the familial cancer syndrome Fanconi anaemia and with an inherited susceptibility to breast cancer. Resulting in a moderate increase in breast cancer risk of around 2-6 fold (Gage et al 2012). In addition, the functions of *PALB2* and its relation to *BRCA2* suggest it to be a likely candidate gene for EOC. Genetic testing for variants in this gene is not currently performed in the genetics clinic, however *PALB2* has been suggested to be the third breast cancer susceptibility gene and deleterious protein-truncating variants appear to affect the risk of developing breast cancer (Rahman et al 2007).

Walsh et al (2011) investigate a group of 12 candidate genes in 326 women with ovarian cancer; the cohort are not enriched for family history or early onset disease and those patients referred for genetic risk are not included in the study. In addition, patients diagnosed with recurrent disease are also excluded. They find that all 12 genes have deleterious variants all of which are in the Fanconi anaemia pathway. Specifically, Walsh et al (2011) detect two deleterious mutations in *PALB2* that are not previously associated with ovarian cancer.

PALB2 is established as a breast cancer susceptibility gene. Rahman et al (2007) examine *PALB2* in breast cancer patients that are screened negative for *BRCA1* or *BRCA2* and discover deleterious mutations in around 1% of patients with familial

breast cancer, which leads to an elevated breast cancer risk of 2.3 x higher than controls. These data are validated by Erkko et al (2007) who discover a deleterious mutation in a Finnish population that results in an increase in breast cancer in familial cases. These deleterious mutations affect the functioning of *BRCA2* and result in inadequate repair by homologous recombination. Erkko (2007) find that this deleterious variant is 4 times more likely to be present in cases compared to healthy age-matched controls. Tischkowitz et al (2012) detect deleterious mutations in 0.9% of breast cancer patients diagnosed with bilateral disease; all of the patients with breast cancer in only one breast are negative for these *PALB2* mutations. Interestingly, this difference is only observed in patients with deleterious frameshift mutations and does not involve missense variants as these occur at the same frequency in both groups. The relative risk of breast cancer for patients with deleterious mutations in *PALB2* is 5.3.

4.2.2 The structure and function of *BRIP1* (BRCA1 interacting protein C-terminal helicase 1)

The gene *BRCA1* interacting protein C-terminal helicase 1 (*BRIP1*) NCBI Genbank accession number NM_032043 and OMIM 605882 is also known as Fanconi anaemia complementation group J (*FANCJ*) or BRCA1-associated C-terminal helicase 1 (*BACH1*). The gene has 20 exons, 19 of which are coding exons; the coding sequence starts in exon 2. The gene is located on chromosome 17q23.2 with genomic co-ordinates chr17:59756547-59940920 and includes a genomic region of 184, 374 bp.

Figure 4.3 Schematic representation of *BRIP1* (BRCA1 interacting protein C-terminal helicase 1) gene

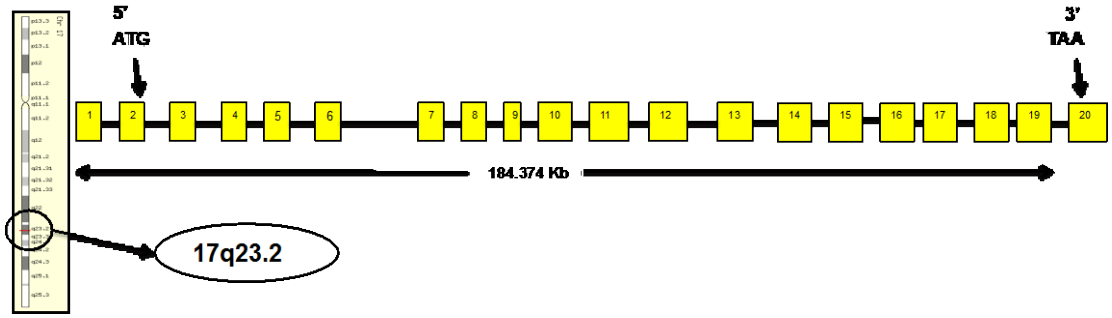


Figure 4.3 Schematic representation of BRIP1. *BRIP1 has 20 exons 19 of which are coding exons and includes a genomic region of 184,374 bp*

The protein product of *BRIP1* is an ATP dependent DNA helicase that forms a complex with BRCA1 protein binding between the two BRCT repeat regions at the C-terminal of

BRCA1 (Cantor et al 2001). It functions in DNA repair directed by BRCA1 and is intimately related to BRCA1 with a role in BRCA1 checkpoint activation. The relationship between *BRIP1* and *BRCA1* is cell-cycle control via the phosphorylation of amino 990 (Ser residue) within *BRIP1* and it is suggested that this interaction assists *BRCA1* in its function in DNA double strand break repair and checkpoint control at the G2/M phase of the cell cycle (Cantor & Guillemette 2011). This gene is shown to be a low-moderate breast cancer susceptibility gene (Cantor & Guillemette 2011). The encoded protein product of *BRIP1* interacts with BRCA1 and germline mutations in *BRIP1* gene affect the function of *BRCA1* in homologous recombination.

4.2.3 The structure and function of *BARD1* (BRCA1-associated RING domain protein 1)

The gene *BARD1* (*BRCA1*-associated RING domain protein 1) NCBI Genbank accession number NM_000465 OMIM 601593 is composed of 11 exons with all 11 as coding exons. It is located on chromosome 2q35 with genomic co-ordinates chr2:215593275-215674428 and a genomic size of 81,154 bp.

Figure 4.4 Schematic representation of *BARD1* (BRCA1-associated RING domain protein 1)

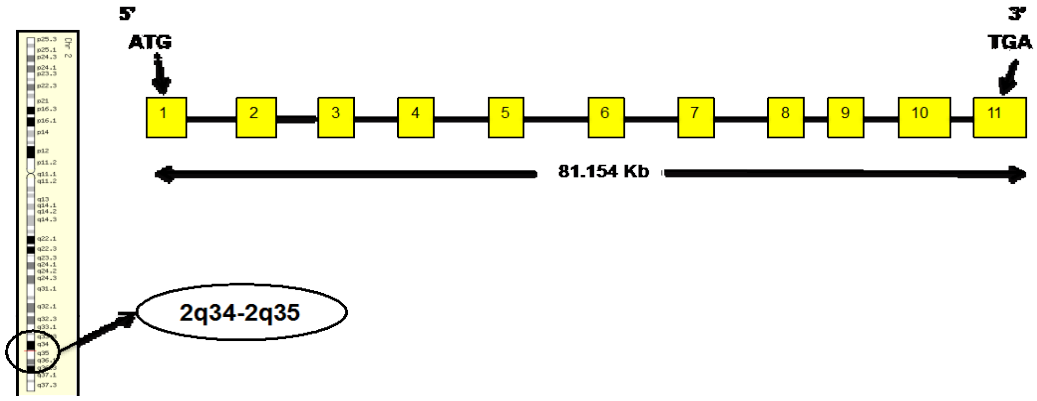


Figure 4.4 Schematic representation of BARD1. *BARD1 has 11 exons all of which are coding exons and includes a genomic region of 81,154 bp*

BARD1 binds the N-terminal of *BRCA1* and possesses homologous regions with two regions in *BRCA1* – the N-terminal RING domain and the C-terminal BRCT repeat domain. A number of proteins involved in cell growth regulation contain RING motifs, which consist of cysteine rich sequences. Studies show that this protein directly affects the function of *BRCA1* and in fact the complex that is formed between *BRCA1* and *BARD1* may be key in the suppression of tumourigenesis (Wu et al 1996). The BARD1

protein is located within the nucleus of the cell and its expression is elevated in cells actively proliferating, including ovary, breast as well as cells of the testes and spleen. When expression of *BRCA1* is reduced this leads to an increase in *BARD1* expression in the cytoplasm and in this state *BARD1* functions in the activation of apoptosis (Ratajska et al 2013). Thus, it follows that non-functioning *BARD1* is very harmful to cells resulting in genomic instability. Ratajska et al (2013) investigate the contribution of germline mutations in *BARD1* in families negative for mutations in either *BRCA1* or *BRCA2*. They discover a number of novel variants including one protein truncating mutation, a splice site variant and a number of predicted deleterious variants.

4.2.4 The structure and function of *Nibrin* (Nijmegen breakage syndrome)

Figure 4.5 Schematic representation of *Nibrin* (Nijmegen breakage syndrome)

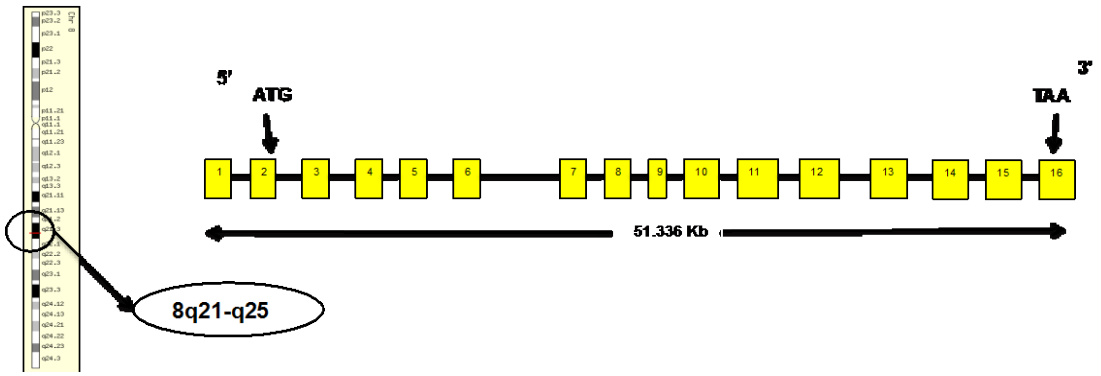


***Figure 4.5 Schematic representation of Nibrin**. Nibrin has 16 exons 15 of which are coding exons and includes a genomic region of 51,336 bp*

Nibrin (Nijmegen breakage syndrome) NCBI accession number NM_002485 OMIM 602667. *Nibrin* has 16 exons, all of which are coding exons. It is located at chromosome 8q21.3 with genomic co-ordinates: chr8:90945564-90996899. The protein product of this gene is involved in DNA double strand break repair as part of a heteropentamer known as MRE11/RAD50 complex (Zhong et al 1999). This complex consists of 5 monomers and includes *MRE11, RAD50, ATM, H2AX, TERF2* and *BRCA1*. It has a role in cell cycle checkpoint activation due to DNA breakage. The Nijmegen breakage syndrome is an autosomal recessive disorder that causes chromosome instability and includes several consequences including susceptibility to various cancers. The MRE11/RAD50 complex interacts with BRCA1 forming nuclear foci with RAD50 (Wang et al 2000).

4.3 Screening for ovarian cancer (CA-125)

Bast et al (1981) report on the discovery of cancer antigen (CA)-125, which can be detected using a murine monoclonal antibody known as OC125. This antibody reacts with six different epithelial ovarian carcinoma cell lines and tumour samples derived from ovarian cancer patients. Bast et al (1981) suggest that a diverse range of normal tissues and normal cell lines do not bind the antibody OC125. Since this discovery much research time has gone into assessing CA-125 and the possibility of using it as a tumour marker and screening tool for ovarian cancer. In addition to CA-125, transvaginal ultrasonography (TVS) made an entrance in an article by Campbell et al (1982). This research evaluates how precisely TVS can quantify ovarian volume and assess ovarian morphology. The researchers suggest that TVS has potential as a novel screening tool for ovarian cancer. Since then TVS has been further refined and is of particular use in postmenopausal women when rates of false positives are substantially decreased. Additionally, performing several sequential TVS scans further eliminates benign conditions of the ovary. Assessment of ultrasound results is improved following clinical screening trials, where follow up of those participants shows that unilocular cysts less than 10 cm in diameter suggest a very low risk of ovarian cancer; in contrast complex cysts not of uniform morphology or that appear to exhibit solid regions suggest a high risk of ovarian cancer. Further, the use of TVS assists in understanding the levels of CA-125 in that an elevated CA-125 result coupled with abnormal ovarian morphology on TVS suggests a high risk of ovarian cancer (Lewis & Menon 2004). Despite these improvements TVS and CA-125 appear to be useful in detecting late stage disease, but still lack the sensitivity to early detect lesions (Long & Kauff 2013). It is the earlier detection of ovarian cancer that is required in order to improve the mortality rates.

4.4 Samples for next generation sequencing

The samples in the study are acquired from a large prospective screening trial conducted in the UK known as the UK Familial Ovarian Cancer Screening Study (UKFOCSS). Screening involves a blood test measuring levels of CA-125 on 3 occasions each year. In addition annual transvaginal ultrasound scans (TVS) are performed. The principal aim of this study is to discover whether screening strategies can early detect ovarian cancers. This trial aims to develop a novel screening approach for women at high risk of developing ovarian cancer due to an assessed genetic susceptibility or a very strong family history and to discover new biomarkers that can diagnose, stage and grade epithelial ovarian cancer and cancer of the

fallopian tube. Women are included into the trial under specific criteria. In families affected by both ovarian and breast cancer the following is applicable:

1. Women FDRs in families where two women or more were affected by ovarian cancer.
2. Women FDRs in families with one ovarian and one breast cancer case diagnosed under 50 years
3. Women FDRs in families with one ovarian and two breast cancer cases under 60 years
4. Breast cancer in the trial member under 45 years old and additionally having a mother with both breast and ovarian cancer.
5. Breast cancer in a trial member under 40 years as well as a sister with both breast and ovarian cancer.
6. Some of these criteria may be changed if there appears to be paternal transmission of inherited predisposition.

The screening study includes families that contain a member with an assessed gene mutation that results in an increased risk of ovarian cancer, for example, *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *PMS1* and *PMS2*. In the case of Lynch Syndrome families, the family has to include more than three members with a Lynch syndrome related cancer and be a FDR of this family with at least one case diagnosed under 50 years old. The Lynch related cancers must involve more than one generation in the family. For families with breast cancer only diagnoses there should be a minimum of 4 cases of breast cancer in the family. If the family contains only 3 breast cancer cases, then one must be under 30 years or all of them under 40 years or include one male breast cancer and one bilateral diagnosis. If the screening volunteer has had breast cancer this must have been diagnosed under the age of 50 years and the individual must have a mother with breast cancer under 30 years or the volunteer is under 30 years and the mother is under 50 at diagnosis. If bilateral breast cancer was diagnosed in the volunteer's mother, then diagnosis should be under the age of 40 (Institute for Women's Health website UKFOCSS eligibility criteria accessed 09-08-2013).

4.4.1 PROMISE 2016

PROMISE 2016 is an acronym for *'Predicting Risk of Ovarian Malignancies, Improved Screening and Early detection*'. It is a large international research effort with an overall

aim to reduce the number of deaths from ovarian cancer by 50% both in the UK and further afield. This research aims to identify women within high-risk groups as well as discovering new technological methods to diagnose ovarian cancer. Thus, PROMISE 2016 aims to predict risk and improve diagnosis of ovarian cancer and to discover the optimum screening method specific to groups of women at risk. Professor Ian Jacobs of University of Manchester is the leading investigator and includes an international collaboration of scientists at UCL, University of Cambridge, University of Southern California and Harvard Medical School.

This study fits with the research ethos of PROMISE 2016 and as such draws on the sample bank of UKFOCSS in order to assist in the aims of the PROMISE study. This study sequences 2,300 women in total.

4.5 Study design

This is a family-based study, which draws upon samples from a large prospective cohort of women whom are part of a UK wide familial ovarian cancer screening trial. The case-control study in Chapter 3 establishes and evaluates a high-throughput sequencing approach, which characterises the penetrance and prevalence of 6 candidate genes. This established technology is used here to sequence 9 known and novel candidate genes in ovarian cancer in women that are disease free at the time of inclusion in the study (PROMISE 2016 samples). The women in this study are pre- and post-menopausal, unaffected and part of families with a history of ovarian cancer and breast cancer, therefore, putting them at high risk of developing the disease.

4.6 Research Aims

The 3 main aims of this study are:

1. To use the established high-throughput DNA sequencing technology to characterise the mutation prevalence in 4 known ovarian cancer susceptibility genes (namely *BRCA1* and *BRCA2* and recently identified genes *RAD51C* and *RAD51D*) in 2,300 unaffected women from high-risk breast-ovarian cancer families.
2. To use the established high-throughput NGS approach to examine the mutation prevalence of 5 novel candidate genes (namely *RAD51B*, PALB2, *NBN*, *BRIP1* and *BARD1*) in the same cohort of 2,300 unaffected women from high-risk breast-ovarian cancer families.

3. To increase throughput to include more genes and evaluate the sequencing coverage.

The overall aim of this study is to discover if women with moderate-high risk mutations are more likely to develop ovarian cancer and if these mutations should be targeted for screening for early disease detection in future follow-up studies.

4.7 Hypotheses under investigation

1. Fluidigm Access Array platform and highly multiplexed Illumina sequencing technologies are scalable methods to enable identification of genetic alleles in 9 known and novel susceptibility genes in ovarian cancer.
2. This method of mutation detection will be able to evaluate the mutation prevalence in 9 known and novel genes in a large series of unaffected women from high-risk breast-ovarian cancer families.
3. This method of mutation detection will be able to identify novel ovarian cancer susceptibility genes.

4.8 Results

(Refer to Chapter 6 Materials and Methods page 279)

4.8.1 Study Design – A prospective family-based study

This prospective cohort study, of women at high-risk of developing ovarian cancer due to family history of breast and/or ovarian cancer, uses the previously established NGS DNA sequencing technology. This research characterises the frequency of 4 genes (*BRCA1*, *BRCA2*, *RAD51C* and *RAD51D*) formerly identified as breast or ovarian cancer susceptibility genes, and 5 (*RAD51B*, *PALB2*, *BRIP1*, *NBN* and *BARD1*) novel candidate ovarian cancer susceptibility genes.

4.9 Target enrichment

Target enrichment and preparation of sequencing libraries is conducted using the Fluidigm Access Array platform. This is performed as previously described in Chapter 3, however the Access Array multiplexing is performed at 10 PCR reactions per well (10 Plex) to enable enrichment and library preparation of the coding regions of all 9 genes on 1 Access Array IFC. The tables in Appendix IX give details of the regions covered in each gene, including the genomic co-ordinates for each amplicon. In Appendix X images are displayed mapping the location of each amplicon in each gene. Where exons are larger than 200bp then overlapping amplicons are designed to ensure that all the coding regions are amplified. The reference assembly used here is the Genome Reference Consortium Human build 37 (GRCh37/hg19) from February 2009. The National Centre for Biotechnology Information (NCBI) accession numbers for each gene are, *BRCA1* (NM_007294.3), *BRCA2* (NM_000059.3), *PALB2* (NM_024675.3), *BRIP1* (NM_032043.2), *BARD1* (NM_000465.2), *NBN* (NM_002485.3), *RAD51B* (NM_133510.3), *RAD51C* (NM_058216.1) and *RAD51D* (NM_002878.3)

Table 4.1 Breakdown of amplicons per gene

Gene	No. Amplicons	% Coding region included
BRCA1	67	100
BRCA2	118	100
RAD51B	24	100
RAD51C	19	100
RAD51D	18	100
PALB2	38	100
BRIP1	45	100
BARD1	25	100
NBN	29	100
Total	384	

Table 4.1 The breakdown of amplicons per gene*. This table details the number of amplicons in the whole experiment. Where there is more than 1 amplicon per exon, amplicons overlap by more than the length of the primer sequence to ensure that all bases are included in the sequencing analysis. Amplicons overlap into the flanking intronic regions by more bases than the length of primer sequences to include all the coding exons and splice sites.*

4.10 Library preparation – quantitation and normalisation of pools

Note: I prepared libraries for 2 sequencing lanes (768 samples and 16 Access Array chips) at Great Ormond Street Molecular Genetics Laboratory. Maria Intermaggio and Andre Kim at University of Southern California prepared the remaining 4 lanes (1,536 samples and 32 Access Array chips)

48 sequencing libraries are prepared on each Fluidigm Access Array chip. This includes the full coding region for all 9 candidate genes i.e. 384 PCR reactions are performed for 48 samples per chip. To reach this 10 PCR reactions are conducted in each well of the Access Array chip. (Note that additional genes are included in these Access Array chips, however, only 9 genes are fully analysed). Each Access Array is pooled to create a pool of 48 barcoded prepared libraries. Then 8 chips are pooled in equimolar quantities to form a pool of 384 individually barcoded libraries. The pools of 384 libraries are run in each flow cell lane on the Illumina HiSeq2000. The whole sequencing study is performed using 6 lanes.

4.10.1 Quantitation of pools

As an initial quality control step, a proportion of the prepared libraries are quantified using Agilent Bioanalyzer DNA 1000 chips with 1µl of each library to confirm that samples have amplified prior to pooling.

4.10.2 Normalisation of pools

Normalisation of pools is performed in an identical manner as described in Chapter 3. For each Illumina flow cell lane, 8 pools of chips are pooled in equimolar quantities and the final concentration calculated to ensure dilution of pools to the correct concentration for the flow cell.

4.10.3 Final concentration

The final concentration required for the flow cell is 10nM in a total volume of 50µl. The dilution factor is calculated from the initial molarity/required final molarity. The volume of DNA to add is calculated by final volume/dilution factor. The appropriate volume of water is added to reach 50µl.

4.10.4 Dilution and final QC for flow cell

Once diluted each final pool was quantified in triplicate using the Agilent Bioanalyzer. The mean is calculated and used as the actual final molarity for pools

4.11 Sequencing Quality Control

4.11.1 Depth of coverage data

For each gene sequenced the percentage of coding bases with read depth greater than 30X is calculated for each sample. Table 4.2 gives the mean percentage of coding bases with greater than 30X coverage for each gene. The table also describes the minimum and maximum percentage of reads sequenced at read depths greater than 30X for each gene.

Table 4.2 Mean and range of percentage of coding bases sequenced at greater than 30X coverage for each gene

Gene	Mean % coding bases sequenced at > 30X depth	Minimum-maximum % coding bases sequenced at >30X depth
BRCA1	94.5	0.1-98.8
BRCA2	93.1	1.0-98.1
PALB2	96.9	0-100
BRIP1	93.6	0-99.3
NBN	93.6	0-100
RAD51B	77.3	0-84.8
RAD51C	90.4	0-96.3
RAD51D	81.6	0-86.2
BARD1	89.5	3.3-95.7

Table 4.2 Mean and range of percentage of coding bases sequenced at greater than 30X coverage for each gene. *This table shows that the read depth by gene is very good with the mean percentage of coding bases sequenced at a read depth exceeding 30X at more than 77% for each gene. The table gives detail on the minimum and maximum percentage of reads that are sequenced at a read depth greater than 30X for each gene.*

Table 4.3 Mean and range of percentage of coding bases sequenced at greater than 30X coverage for each gene for 2,200 samples

Gene	Mean % coding bases sequenced at > 30X depth	Minimum-maximum % coding bases sequenced at >30X depth
BRCA1	96.1	78.5-98.8
BRCA2	95.2	71.9-98.1
PALB2	98.8	78.1-100
BRIP1	96.1	53.2-99.3
NBN	96.5	36-100
RAD51B	79.2	43.7-84.8
RAD51C	92.9	47.9-96.3
RAD51D	82.9	55.2-86.2
BARD1	91.5	48.7-95.7

Table 4.3 Mean and range of percentage of coding bases sequenced at greater than 30X coverage for each gene for 2,200 samples*. This table describes the mean and maximum and minimum percentage of coding bases sequenced at a read depth greater than 30X for 2,200 samples. The 100 poorly performing samples are removed here and this demonstrates the good coverage statistics for the study. The best coverage statistics are for genes BRCA1, BRCA2 and PALB2 where 70-100% of coding bases are sequenced at read depths exceeding 30X for 2,200 samples.*

The range for meeting the minimum read depth of 30X for each gene varies from 0.1% and 100%. Analysis of the coverage data in greater detail shows a small subset of samples do not perform well. If the 100 poorly performing samples are removed then all of the remaining 2,200 samples are sequenced with a read depth more than 30X for 80% of the coding bases in *BRCA1* and *BRCA2* together. Table 4.3 gives the mean percentage of coding bases with greater than 30X coverage and the minimum and

maximum percentage of coding bases sequenced at greater than 30X for all genes if these 100 poorly performing samples are removed. This table shows that sequencing performance is good for all genes, especially *BRCA1*, *BRCA2* and *PALB2* in which 70-100% of coding bases are sequenced at read depths greater than 30X.

4.11.2 Filtering out technical artefacts

Reads are filtered based on a minimum read depth of 30X. Following this Sanger validation is conducted on those variants detected where the alternate allele frequencies are ≥ 30%.

4.11.3 Sanger sequencing validation

Maria Intermaggio and Dr Susan Ramus at USC conduct Sanger sequencing validation on all protein-truncating and predicted protein-truncating variants detected by NGS. 113 variants are Sanger sequenced where alternate allele frequencies are at least 30%; of these 89 are validated. Then further Sanger sequencing validation is conducted on those variants where the depth is at least 30X, but the alternate allele frequency is between 20% and 30% and the quality scores are 99; one additional variant is Sanger sequenced in which the quality score is 99, depth 30X and alternate allele frequency is 10%.

For those variants with a read depth of <30X Sanger sequencing is performed on all variants with quality scores of 99 and a read depth above 10X. One variant is the same as another variant, but with a quality score of less than 99 (81). Of these 7 variants are Sanger sequenced and 14% of them are validated. Images of Sanger sequencing results are given following the tables describing detailed results for protein-truncating and predicted protein-truncating variants detected in each gene. This results in 116 variants in total being Sanger sequenced and 99 (85.4%) variants are validated. This means that there are 17 false positive results. There is 1 Sanger sequencing trace representing 1 example variant for each gene (Figures 4.6, 4.8, 4.9, 4.10, 4.11 and 4.12). Further Sanger sequencing validation is conducted following the 2nd analysis, however these data are not available at the time of writing.

4.12 Genetic variant prevalence and characteristics following 1st analysis

This sequencing analysis focuses on those variants predicted to result in protein-truncation; these include, frameshift insertions, deletions and substitutions, nonsense, splice site and functional missense variants. Many of the variants detected in *BRCA1* and *BRCA2* are known and are included in the Breast Cancer Information Core (BIC) database. The analysis is performed twice for two reasons: 1) the first analysis does not include an analysis of splice site changes and 2) the first analysis shows discordance with clinical testing results and therefore, the filtering parameters are relaxed in the 2nd analysis in order to detect the missing variants. Table 4.4 is a summary of the genetic variants according to variant type detected at the 1st analysis.

Table 4.4 Summary of genetic variants detected by gene and variant type at the 1st analysis

Gene	Nonsense	Frameshift deletion	Frameshift insertion	Percentage of protein-truncating variants
BRCA1	4	33	0	1.6
BRCA2	3	37	3	1.9
BRIP1	1	4	0	0.22
PALB2	0	4	0	0.17
NBN	0	3	0	0.13
BARD1	0	1	0	0.04
RAD51C	0	3	0	0.13
RAD51D	0	1	0	0.04

Table 4.4 Summary of genetic variants detected by gene and variant type at the 1st analysis*. This table describes the protein truncating and predicted protein-truncating variants are found in 8 out of 9 genes in the study. NB. Although 99 variants are validated, 2 of these are non-frameshift deletions and are therefore, not included in this table.*

Table 4.4 shows that 1.6 % (n=37) of samples are found to have a protein-truncating variant in *BRCA1*; 1.9% (n=43) of samples are found to have a protein-truncating variant in *BRCA2*. 0.73% of samples have a protein-truncating variant in one of the other genes, with around half of these being in *BRIP1* (0.22%) and *PALB2* (0.17%).

4.12.1 Discordance with clinical testing results and splice site analysis

Comparison with clinical testing results reveals that several mutations are not detected in the initial analysis with the filtering parameters used. The stringency of these filters means that some variants are excluded; however, these cannot be considered false

negatives as they are in the Illumina data analysis when the filtering stringency is relaxed. These variants are validated using Sanger sequencing.

Using clinical data for the trial volunteers whom previously tested positive (n=73) for mutations in either *BRCA1* or *BRCA2*, 14 of these were missed (9 *BRCA1* and 5 *BRCA2*) in the first analysis. NGS data for these are re-investigated and all mutations are detected in NGS data except one variant, 185delAG, which is identified as a SNP in exon 2 rather than a 2bp deletion. One sample is missed due to low coverage in the position of the mutation and the additional 12 are missed due to filtering of the variant annotations using the Genome Analysis Toolkit (GATK) set of analysis tools.

Variant annotation filters from GATK that result in missing these 12 mutations include the following calculations; *Haplotype Score*, *Read Pos Rank Sum Test*, *In breeding Coefficient*, *MQ*, *QD* and *MQ* Rank Sum Test. 11/12 samples are missed due to high haplotype scores, which would suggest regions where alignments are poor leading to sequencing artefacts and false SNP or indel calls. In this instance the use of this filter in variant annotation results in missing 11 mutations in the sequencing data. The *Read Pos Rank Sum Test* makes a calculation of the distance the variant is located from the end of reads in the alternate allele. If these alternate alleles are located repetitively at the end of reads this is filtered out as an artefact. 1/12 is missed solely due to low *QD* (low Quality Depth by Depth). This filter calculates quality divided by unfiltered read depth and a low score suggests a false positive or sequencing artefact. The *Inbreeding Coefficient* is an algorithm based on the Hardy-Weinberg test that suggests that in a randomly breeding population genotype frequency maintains equilibrium. The *MQ* score calculates the mapping quality of variants called, filtering variants out if this score is low. The *MQ* score is a comparison of the quality of reads that contain the reference allele and the quality of reads containing the alternate allele (Mckenna et al 2010, DePristo et al 2011). The initial analysis does not include variants residing in splice sites. In the 2nd analysis 3 splice site changes are detected (Table 4.5 Final deleterious variants detected in candidate genes).

4.13 Genetic variant prevalence and characteristics following 2nd analysis

Table 4.5 Final summary of predicted deleterious variants detected by gene and variant type at the 2nd analysis

Gene	Nonsense	Frameshift deletion	Frameshift insertion	Frameshift substitution	Missense	Splice site	S-SNV
BRCA1	7	41	0	2	2	0	1
BRCA2	7	37	3	0	1	1	0
BRIP1	1	4	0	0	0	0	0
PALB2	0	4	0	0	0	0	0
NBN	0	3	0	0	0	2	0
BARD1	0	1	0	0	0	0	0
RAD51C	0	3	0	0	0	0	0
RAD51D	0	1	0	0	0	0	0

Table 4.5 Final summary of predicted deleterious variants detected by gene and variant type at the 2nd analysis. *NGS identifies 121 samples with predicted deleterious gene variants in 8 of the 9 genes. Sanger sequencing validation is performed on all 121 samples, however data from this validation are not available at the time of writing (September 2013).*

Table 4.5 shows that 2.26 % (n=52) of samples are found to have a predicted deleterious variant in *BRCA1*; 2.13% (n=49) of samples are found to have a predicted deleterious variant in *BRCA2*. In 0.82% of samples, a predicted deleterious variant is detected in one of the other genes, *BRIP1* (0.22%, n=5), *PALB2* (0.17%, n=4) and NBN (0.22%, n=5) and *RAD51C* (0.13% n=3); *RAD51D* and *BARD1* (both 0.04%, n=1). As the women here are unaffected and with a strong family history it could be argued that potentially the real prevalence of mutations might be doubled in affected cases. This could be inferred if the inheritance pattern of these variants is autosomal dominant, in which case subjects would have a 50% chance of inheriting the mutation. Certainly, this is the case with *BRCA1* and *BRCA2* genes in Hereditary Breast Ovarian Cancer Syndrome.

Table 4.6 Detailed results table of predicted deleterious variants detected in *BRCA1*

Gene	Function	BIC	cDNA	Protein	Exon	Frequency	On BIC Y/N
BRCA1	FS Del	5083del19-ter1658	c.4964_4982del19	p.Ser1655fs	15	1	Y
	FS Del	4491_4507del17(STOP1469)	c.4372_4388del17	p.Gln1458fs	13	1	Y
	FS Del	4232delG	c.4113delG	p.Cys1372fs	11	1	Y
	FS Del	4184_4187delTCAA(STOP1364)	c.4065_4068delTCAA	p.Asn1355fs	11	9	Y
	FS Del	3875del4(STOP1262)	c.3756_3759delGTCT	p.Leu1252fs	11	5	Y
	FS Del	3448del4(STOP1115)	c.3329_3332delAGCA	p.Lys1110fs	11	1	Y
	FS Del	3121delA(STOP1023)	c.3002_3002delA	p.Glu1001fs	11	1	Y
	FS Del	3006delA(STOP999)	c.2887delA	p.Thr963fs	11	1	Y
	FS Del	2823delG	c.2704delG	p.Glu902fs	11	1	N
	FS Del	2776_2777delCT	c.2657_2658delCT	p.Ser886fs	11	2	Y
	FS Del	2246delT	c.2127delT	p.Phe709fs	11	1	N
	FS Del	1624_1628del4	c.1505_1509delTAAAG	p.Leu502fs	11	4	N
	NS	1445T>A	c.1326T>A	p.Cys442X	11	1	N
	FS Del	917delTT-ter285	c.798_799delTT	p.Ser267fs	10	1	Y
	FS Del	795delT-ter233	c.676delT	p.Cys226fs	10	2	Y
	NS	546G>T(E143X)	c.427G>T	p.Glu143X	6	3	Y
	FS Del	221delT	c.102delT	p.Pro34fs	2	1	N
	FS Del	185delAG	c.67_67delinsAG	p.Glu23Val	2	7	Y
	FS Del	3274delA	c.3155delA	p.Asn1052fs	11	1	N
	NS	E1134X	c.3400G>T	p.Glu1134X	11	1	Y
	NS-SNV	L204F	c.612G>C	p.Leu204Phe	10	1	Y
	NS-SNV	M1775R	c.5324T>G	p.Met1775Arg	21	1	Y
	FS Sub	2080insA	c.1960_1962AAAG	p.Lys654Lys	11	1	Y
	NS	W1718X	c.5153G>A	p.Trp1718X	18	2	Y
	S-SNV	Q1395Q	c.4185G>A	p.Gln1395=	12	1	Y
	FS Sub	5386delinsCA	c.5267_5267delinsCA	pGln1756ProfxX74	20	1	N

Table 4.6 Detailed results table of predicted deleterious variants detected in BRCA1. *This table describes details of predicted deleterious variants detected in BRCA1, including mutation type, frequency, position and if known on BIC database. NS = nonsense (Stop Gain), FS Del = Frameshift deletion, FS Sub = Frameshift substitution, NS-SNV = non-synonymous single nucleotide variant, S-SNV = synonymous single nucleotide variant. 7 mutations were not previously recorded on the BIC database.*

Table 4.6 gives detailed results for NGS in *BRCA1* gene. 26 different variants are detected in *BRCA1* and 7 of these are not listed on the BIC database or on the NCBI dbSNP database; of these 5 are unknown frameshift deletions, 1 is an unknown frameshift substitution and 1 is an unknown nonsense variant. Without performing functional analyses these can only be termed predicted deleterious variants and

therefore are variants of uncertain significance. Variants of uncertain significance may or may not be clinically relevant.

Figure 4.6 Sanger sequencing validation image of *BRCA1* protein-truncating mutation. Frameshift deletion c.1505_1509delTAAAG

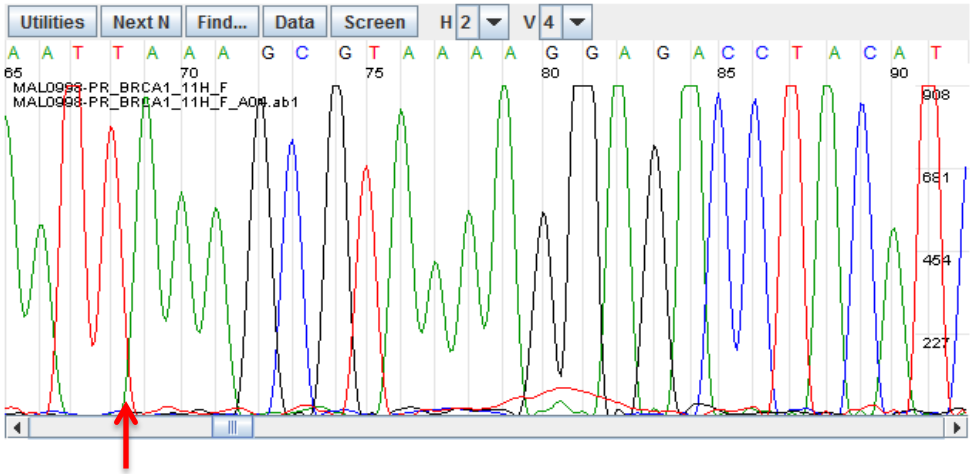


***Figure 4.6 Sanger sequencing validation image of BRCA1 protein-truncating mutation**. This image is a Sanger sequencing trace showing the frameshift deletion c.1505_1509delTAAAG. The red arrow indicates deletion of the 5 nucleotides TAAAG in BRCA.*

All samples with NGS identified protein truncating and predicted protein truncating variants in *BRCA1* are Sanger sequenced for validation purposes. Figure 4.6 is an example of 1 Sanger sequencing validation image for the frameshift deletion c.1506_1509delTAAAG.

Figure 4.7 Read alignment of read with mutation185delAG

```
00000001 acactaagattttctgcatagcaataatgacattttgtacttcttcnacg 00000050  SEQUENCE
>>>>>>>> || |||||||||||||||||||| |||||||||||||||||||||||| ||| >>>>>>>>
41276044 actctaagattttctgcatagcattaatgacattttgtacttcttcaacg 41276093  REFERENCE
```

If the read had been longer, the actual alignment might look like this:

```
00000001 atgggaca..ctaagattttctgcatagcaataatgacattttgtacttc 00000048  SEQUENCE READ
>>>>>>>> ||||||||  |||||||||||||||||||| |||||||||||||||||||| >>>>>>>>
41276037 atgggacactctaagattttctgcatagcattaatgacattttgtacttc 41276086  REFERENCE
```

***Figure 4.7 Read alignment of read with mutation 185delAG.** The top read alignment is the actual one, the 3rd base from the 5' end shows a SNP A>T where the mutation should be on the reverse sequence i.e. deletion of CT in forward sequence. By simulating an increase in read length it is possible to see the deletion of the 2 bases (in the bottom sequence).*

1 frameshift deletion 185delAG is initially detected by NGS as a SNP not a deletion and the reason for this is that the read containing this mutation is misaligned by the aligner script because the mutation is located at the end of a read. If reads are longer or if there is enough overlap of amplicons then misalignments should not happen. The image (Figure 4.7) above describes this problem by simulating a longer read in the second example.

Table 4.7 Detailed results table of predicted deleterious variants detected in *BRCA2*

Gene	Function	BIC	cDNA	Protein	Exon	Frequency	On BIC Y/N
BRCA2	FS Del	248_249delAG	c.20_21delAG	p.Glu7fs	2	1	N
	FS Del	635delA	c.407delA	p.Asn136fs	4	1	Y
	FS Del	0886delGT-ter220	c.658_659delGT	p.Val220fs	8	1	Y
	FS Del	982del4	c.755_758delACAG	p.Asp252fs	9	3	Y
	FS Del	1493delA	c.1265delA	p.Asn422fs	10	1	Y
	NS	1684C>T (Q486X)	c.1456C>T	p.Gln486X	10	1	Y
	FS Del	2157delG-ter659	c.1929delG	p.Arg645fs	11	1	Y
	FS Del	3773delTT-ter1182	c.3545_3546delTT	p.Phe1182X	11	1	Y
	FS Del	3908delTG-ter1231	c.3680_3681delTG	p.Leu1227fs	11	1	Y
	FS Del	3917delC	c.3689delC	p.Ser1230fs	11	1	Y
	FS Del	4075_4076delGT(STOP1284)	c.3847_3848delGT	p.Val1283fs	11	2	Y
	FS Del	4626del5	c.4398_4402del5	p.Leu1466fs	11	2	Y
	FS Del	4638del4	c.4410_4413delAAGA	p.Ile1470fs	11	1	N
	FS Del	4705del4	c.4477_4480delGAAA	p.Glu1493fs	11	1	Y
	FS Del	5104delAA	c.4876_4877delAA	p.Asn1626fs	11	2	Y
	FS Ins	5294_5295insA	c.5066_5067insA	p.Ala1689fs	11	3	Y
	NS	5507C>G-Ser1760ter	c.5279C>G	p.Ser1760X	11	1	Y
	FS Del	5531delTT-ter1772	c.5303_5304delTT	p.Leu1768fs	11	3	Y
	FS Del	5578delAA(STOP1785)	c.5350_5351delAA	p.Asn1784fs	11	1	Y
	FS Del	6174delT	c.5946delT	p.Ser1982fs	11	4	Y
	FS Del	6503delTT-ter2098	c.6275_6276delTT	p.Leu2092fs	11	5	Y
	FS Del	6926delC	c.6698delC	p.Ala2233fs	11	1	N
	FS Del	7297delCT-ter2358	c.7069_7070delCT	p.Leu2357fs	14	2	Y
	FS Del	8803delC-ter2862	c.8575delC	p.Gln2859fs	20	1	Y
	NS	9610C>T(R3128X)	c.9382C>T	p.Arg3128X	25	1	Y
	FS Del	2116delA	c.1888delA	p.Thr630fs	10	1	N
	NS	W2586X	c.7757G>A	p.Trp2586X	16	1	Y
	NS	W563X	c.1689G>A	p.Trp563Ter	10	2	Y
	NS	S1970X	c.5909C>A	p.Ser1970Ter	11	1	Y
	NS-SNV	E2663V	c.7988A>T	p.Glu2663Val	18	1	Y
	Splice	8875delG	c.8756delG	P.Gly2919Valfs	Intronic	1	Y

Table 4.7 Detailed results table of predicted deleterious variants detected in BRCA2*. This table describes details of predicted deleterious variants detected and validated in BRCA2, including mutation type, frequency, position and if known on BIC database. NS = nonsense (Stop Gain), FS Del = Frameshift deletion, NS-SNV = non-synonymous single nucleotide variant. 4 variants were not previously recorded on the BIC database*

Table 4.7 gives detailed results for NGS in *BRCA2* gene. 31 different variants are detected in *BRCA2* with 4 of these not listed on the BIC database or on the NCBI dbSNP database; of these all 4 are unknown frameshift deletions. Without performing functional analyses these can only be termed predicted deleterious variants and therefore are variants of uncertain significance with untested clinical relevance.

Figure 4.8 Sanger sequencing validation image of *BRCA2* protein-truncating mutation. Nonsense mutation in *BRCA2* c.1456C>T Q486X

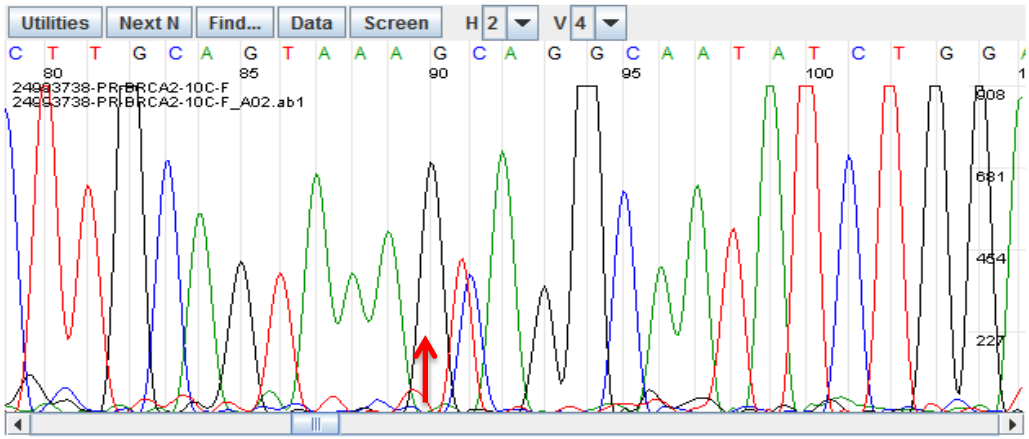


***Figure 4.8 Sanger sequencing validation image of BRCA2 protein-truncating mutation**. This image is a Sanger sequencing trace showing the nonsense mutation c.1456C>T, the red arrow indicates the single base change C>T in exon 10 of BRCA2 resulting in insertion of stop codon*

All samples with NGS identified variants in *BRCA2* are Sanger sequenced for validation purposes. The image in Figure 4.8 is one example of a *BRCA2* protein-truncating mutation detected in exon 10 of the gene.

Table 4.8 Detailed results table of protein-truncating mutations detected in *BRIP1*

Gene	Function	cDNA	Protein	Exon	Frequency	Novel or Known
BRIP1	FS Del	c.2990_2993delCAAA	p.Thr997fs	20	1	Novel
	FS Del	c.2255_2256delAA	p.Lys752fs	15	1	Novel
	FS Del	c.890delA	p.Lys297fs	7	1	Novel
	FS Del	c.128_131delTGTT	p.Leu43fs	3	1	Novel
	NS	c.66C>A	p.Tyr22X	2	1	Novel

***Table 4.8 Detailed results table of protein-truncating mutations detected in BRIP1.** This table describes detail of protein-truncating mutations detected in BRIP1, including mutation type, frequency, position and if previously identified in literature (novel or known).*

Table 4.8 gives detailed results of predicted protein-truncating variants in *BRIP1*. There are 4 different frameshift deletions and 1 nonsense variant. All of these only occur once and all are novel variants not on dbSNP or previously published.

Figure 4.9 Sanger sequencing validation image of *BRIP1* protein-truncating mutation. Nonsense mutation in exon 2 BRIP1 c.66C>A

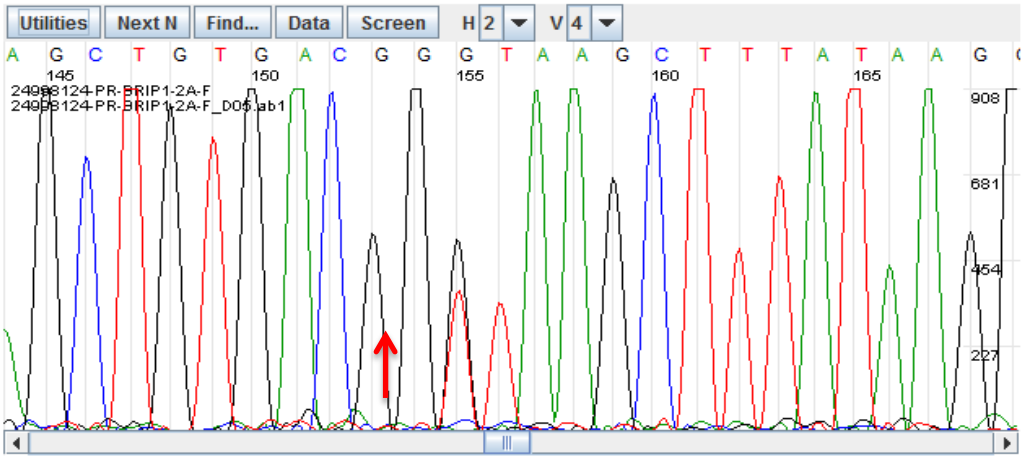


***Figure 4.9 Sanger sequencing validation image of BRIP1 protein-truncating mutation**. This image is a Sanger sequencing trace showing the nonsense mutation in exon 2 of BRIP1 c.66C>A. The red arrow indicates single base change C>A in exon 2 of BRIP1 resulting in insertion of stop codon*

All samples with *BRIP1* predicted protein-truncating variants are Sanger sequenced for validation purposes. The image in Figure 4.9 is one example of a *BRIP1* predicted protein-truncating mutation detected in exon 2 of the gene.

Table 4.9 Detailed results table of protein-truncating mutations detected in *PALB2*

Gene	Function	cDNA	Protein	Exon	No. Samples	Novel or Known
PALB2	FS Del	c.2488delG	p.Glu830fs	5	1	Novel
	FS Del	c.2167_2168delAT	p.Met723fs	5	1	Novel
	FS Del	c.509_510delGA	p.Arg170fs	4	1	rs515726124 known pathogenic
	FS Del	c.172_175delTTGT	p.Leu58fs	3	1	rs180177143 known pathogenic

***Table 4.9 Detailed results table of protein-truncating mutations detected in PALB2.** This table describes details of protein truncating mutations detected in PALB2, including mutation type, frequency, position and if previously identified in literature (novel or known).*

Table 4.9 gives a detailed description of predicted protein-truncating variants detected in *PALB2*. 4 frameshift deletions are detected with 2 being novel and not previously reported in literature or listed in dbSNP 2 frameshift deletions listed on dbSNP. Both variants c.509_510delGA and c.172_175delTTGT are listed in dbSNP as short genetic variants that are known pathogenic and therefore are not given minor allele frequencies. Casadei et al (2011) published both of these mutations. They report 13 different protein-truncating mutations in the gene in breast cancer patients with family history of the disease.

Figure 4.10 Sanger sequencing validation image of *PALB2* protein-truncating mutation. Frameshift deletion in exon 5 of *PALB2* c.2488delG

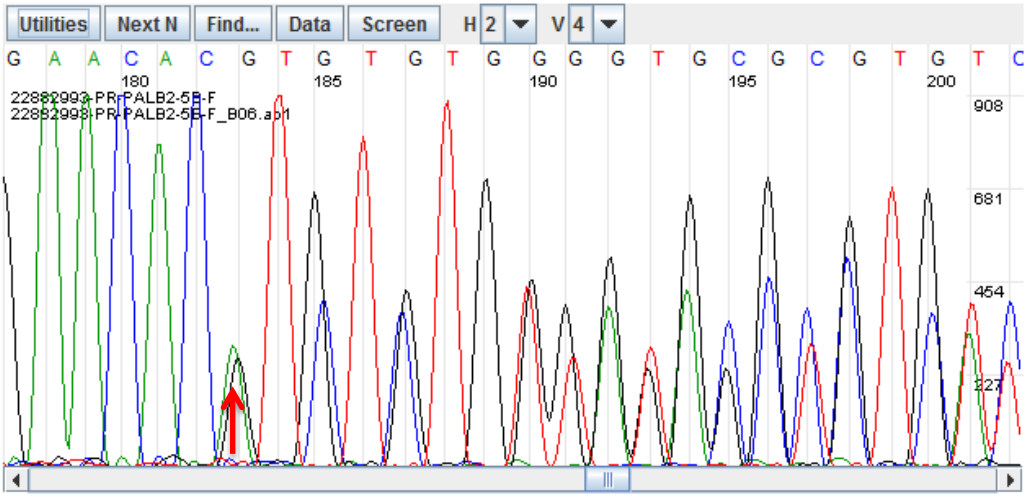


Figure 4.10 Sanger sequencing validation image of PALB2 protein-truncating mutation. Frameshift deletion in exon 5 of PALB2 c.2488delG. *The red arrow indicates single base deletion of G in exon 5 of PALB2 resulting in a frameshift mutation*

All samples with *PALB2* predicted protein-truncating variants are Sanger sequenced for validation purposes. The image in Figure 4.10 is one example of a novel *PALB2* predicted protein-truncating mutation detected in exon 5 of the gene.

Table 4.10 Detailed results table of predicted protein-truncating mutations detected in *NBN*

Gene	Function	cDNA	Protein	Exon	No. Samples	Novel or Known
NBN	FS Del	c.1142delC	p.Pro381fs	10	2	Novel
	FS Del	c.657_661delACAAA	p.Lys219fs	6	1	Novel
	Splice	c.481-1G>A		6	2	Novel

Table 4.10 Detailed results table of protein truncating mutations detected in NBN. *This table describes details of protein truncating mutations detected in NBN, including mutation type, frequency, position and if previously identified in literature (novel or known).*

Table 4.10 details the predicted protein-truncating variants detected in *NBN* gene. 3 different predicted protein-truncating variants (2 frameshift deletions and 1 splice site) are detected in 5 samples. All variants are novel and not reported in literature or listed in dbSNP.

Figure 4.11 Sanger sequencing validation image of *NBN* predicted protein-truncating variant. Frameshift deletion in exon 10 of *NBN* c.1142delC

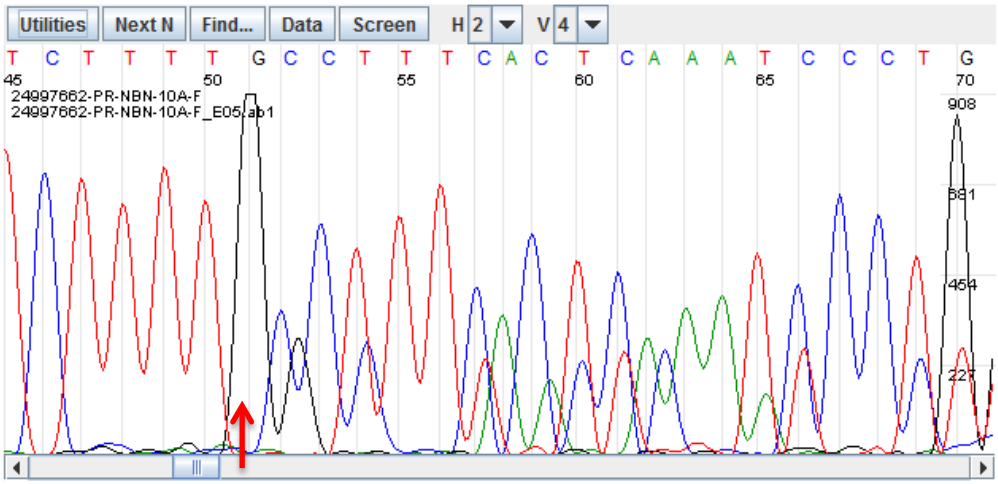


Figure 4.11 Sanger sequencing validation image of NBN predicted protein-truncating variant. Frameshift deletion in exon 10 of NBN c.1142delC. *The red arrow indicates single base deletion of C in exon 10 of NBN resulting in a frameshift mutation*

All samples with *NBN* predicted protein-truncating variants are Sanger sequenced for validation purposes. The image in Figure 4.11 shows the frameshift deletion in exon 10 of the gene.

Table 4.11 Detailed results table of predicted protein truncating variants detected in *RAD51C*, *RAD51D* and *BARD1*

Gene	Function	cDNA	Protein	Exon	No. Samples	Novel or Known
RAD51C	FS Del	c.158delC	p.Ser53fs	2	1	Novel
RAD51C	FS Del	c.731delT	p.Ile244fs	5	2	Novel
RAD51D	FS Del	c.748delC;c.808delC;c.412delC	p.His250fs;p.His270fs;p.His138fs	9,9,6	1	Novel
BARD1	FS Del	c.2291_2294delTAGA	p.Ile764fs	11	1	Novel

Table 4.11 Detailed results table of protein truncating mutations detected in RAD51C, RAD51D and BARD1. *This table describes details of protein truncating mutations detected in RAD51C, RAD51D and BARD1 including mutation type, frequency, position and if previously identified in literature (novel or known).*

Table 4.11 gives detailed results of predicted protein-truncating variants detected in *RAD51C*, *RAD51D* and *BARD1*. There are 2 different frameshift deletions in *RAD51C* in 3 samples and 1 frameshift deletion in each of *RAD51D* and *BARD1* in 1 sample in each gene. All of the variants in these 3 genes are novel and not previously published or listed in dbSNP.

Figure 4.12 Sanger sequencing validation image of *BARD1* predicted protein-truncating variant. Frameshift deletion in exon 11 c.2291_2294delTAGA

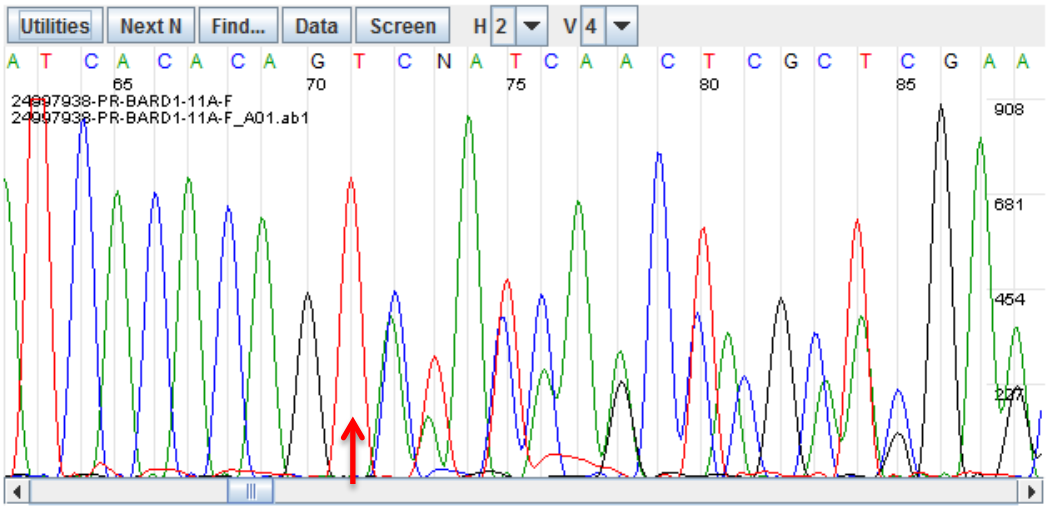


***Figure 4.12 Sanger sequencing validation image of BARD1 predicted protein-truncating variant. Frameshift deletion in exon 11 c.2291_2294delTAGA**. The red arrow indicates 4 base deletion in exon 11 of BARD1 resulting in a frameshift mutation*

All samples with predicted protein-truncating variants in *BARD1*, *RAC51C* and *RAD51D* are Sanger sequenced for validation purposes. The image in Figure 4.12 shows the Sanger sequencing trace for the frameshift deletion in exon 11 of *BARD1*.

4.14 Epidemiological data

Epidemiological data are available for 94 of the 121 samples with protein-truncating or predicted protein-truncating variants in each gene. The tables 4.12 to 4.14 give detail on family history for each sample with an NGS detected variant and gives age at recruitment to the study. This family history data includes the number of affected relatives and the type of cancer diagnosed. If the volunteer has had a diagnosis of cancer then the age of diagnosis and cancer type are specified. If the volunteer has been tested in an NHS clinic for *BRCA1* or *BRCA2* then the result of this test is given.

Table 4.12 Epidemiological data for samples positive for *BRCA1* gene variants

Sample	Variant type	Variant BIC	Epidemiological data				
			Family History	Age at recruitment	Cancer Dx Y/N	Age Cancer Dx	NHS *BRCA1/2*
1	FS Del	5083del19-ter1658	2 ovarian 2 breast 1 stomach 1 leukaemia	54	No	Na	Not tested
2	FS Del	4491_4507del17 (STOP1469)	4 breast 1 ovarian 2 colon 1 lung	52	no	Na	Not tested
3	FS Del	4232delG	2 breast 1 brain 1 stomach 1 uterus 1 lung 2 ovarian	37	breast cancer	31	Not tested
4	FS Del	4184_4187del TCAA (STOP1364)	3 ovarian 2 breast	39	no	Na	*BRCA1* positive
5	FS Del	4184_4187del TCAA (STOP1364)	3 breast 2 ovarian 1 leukaemia	54	no	Na	Not tested
6	FS Del	4184_4187del TCAA (STOP1364)	4 ovarian 6 breast 4 brain 1 unknown primary	44	no	Na	Not tested
7	FS Del	4184_4187del TCAA (STOP1364)	3 breast 1ovarian 1 leukaemia 1 oesophageal 1 prostate	44	No	Na	Not tested
8	FS Del	4184_4187del TCAA (STOP1364)	2 ovarian 1 breast 1 unknown primary	42	No	Na	Not tested

9	FS Del	4184_4187del TCAA (STOP1364)	2 breast 1 ovarian	42	No	No	Not tested
10	FS Del	4184_4187del TCAA (STOP1364)	5 Liver 1 ovarian 3 breast 1 oesophageal 1 unknown primary	39	Breast	40	*BRCA1* positive
11	FS Del	4184_4187del TCAA (STOP1364)	2 breast 1 ovarian 1 endometrial 1 stomach 1 prostate	51	No	Na	Not tested
12	FS Del	4184_4187del TCAA (STOP1364)	3 ovarian 1 alimentary tract	46	No	Na	Not tested
13	FS Del	3875del4 (STOP1262)	3 ovarian 3 breast 1 endometrial 1 oropharynx 1 lung/bronchial 1 alimentary tract	43	Breast cancer	40	*BRCA1* positive
14	FS Del	3875del4 (STOP1262)	2 ovarian 1 colorectal 1 unknown primary	41	No	Na	Not tested
15	FS Del	3875del4 (STOP1262)	1 ovarian 2 alimentary tract 2 unknown primary	39	Breast cancer	29	Not tested
16	FS Del	3875del4 (STOP1262)	4 breast 1 ovarian 1 colorectal 1 lung 1 gastric	38	Cervical cancer	40	Not tested
17	FS Del	3875del4 (STOP1262)	5 ovarian 1 peritoneal 1 breast 1 gastric 1 unknown primary	60	No	Na	*BRCA1* positive
18	FS Del	3448del4 (STOP1115)	1 ovarian 1 breast 1 prostate 2 thyroid 1 malignant melanoma 1 unknown primary	44	No	Na	Not tested

19	FS Del	3121delA (STOP1023)	1 breast and ovarian 1 breast	42	No	Na	Not tested
20	FS Del	3006delA (STOP999)	2 breast 1 pancreatic 1 Hodgkin's lymphoma 1 liver 1 parotid gland 2 skin and nails 1 neck and spine	52	Breast cancer	41	Not tested
21	FS Del	2823delG	5 breast 1 ovarian	42	No	Na	Not tested
22	FS Del	2776_2777delCT	2 ovarian 1 cervical	41	No	Na	Not tested
23	FS Del	2776_2777delCT	1 ovarian 1 breast 1 osteosarcoma 1 unknown primary	38	No	Na	Not tested
24	FS Del	2246delT	2 ovarian 7 breast 1 neck and spine 2 malignant melanoma 2 lung/bronchial 1 carcinoid bronchus 1 heart 1 testicular 2 bladder 2 Hodgkin's lymphoma 2 endometrial 1 leukaemia	46	Breast cancer	47	*BRCA1* positive
25	FS Del	1624_1628del4	1 breast 1 ovarian	54	Breast cancer	39	*BRCA1* positive
26	FS Del	1624_1628del4	2 ovarian 2 breast	46	No	Na	*BRCA1* positive
27	FS Del	1624_1628del4	1 ovarian 1 breast 1 leukaemia 1 cervical	46	No	Na	Not tested
28	Nonsense	1445T>A	3 ovarian 3 breast	43	Ovarian	45	*BRCA1* positive
29	FS Del	917delTT-ter285	1 ovarian 4 breast 1 colorectal 1 cervical	39	No	Na	Not tested
30	FS Del	795delT-ter233	3 ovarian 1 cervical 1 lung/bronchial 2 alimentary	36	No	Na	Not tested

31	FS Del	795delT-ter233	3 ovarian 2 endometrial 2 liver 4 gastric	38	No	Na	Not tested
32	Nonsense	546G>T(E143X)	3 ovarian 1 breast 1 pancreatic 1 kidney 1 lung/bronchial 4 unknown primary	60	No	Na	*BRCA1* positive
33	Nonsense	546G>T(E143X)	1 ovarian 3 breast	49	No	Na	Not tested
34	Nonsense	546G>T(E143X)	1 ovarian 6 breast 1 uterus 2 prostate 1 pancreatic 2 malignant melanoma	38	No	Na	Not tested
35	FS Del	221delT	3 ovarian 1 breast	36	No	Na	Not tested
36	FS Del	3274delA	1 ovarian 1 breast 1 colorectal 1 lung/bronchial	38	No	Na	Not tested

Table 4.12 Epidemiological data for samples positive for BRCA1 gene variants*. This table gives details on family history for 36 samples with protein-truncating variants in BRCA1 gene. 5 of these had prior breast cancer diagnoses; 4 are subsequently diagnosed with cancer (2 breast cancer, 1 ovarian cancer and 1 cervical cancer) fairly soon after joining the study. The remaining 27 samples do not report cancer diagnoses at the time of writing (July 2013). Dx = diagnosis*

Table 4.12 shows epidemiological data for 36 women positive for protein-truncating variants in *BRCA1*. These data show that 27 women with *BRCA1* variants are currently unaffected by breast or ovarian cancer and all of whom have a strong family history of various cancer types including breast and/or ovarian. The ages of women range from 36 years to 60 years at recruitment, therefore, representing pre-, peri- and post-menopausal women; of the women with prior or subsequent cancer diagnoses the age of diagnosis ranges from 29 years to 47 years.

Table 4.13 Epidemiological data for samples positive for *BRCA2* gene variants

Sample	Variant Type	Variant BIC	Epidemiological Data				
			Family History	Age at recruitment	Cancer Dx Y/N	Age Cancer Dx	NHS *BRCA1/2*
37	FS Del	248_249delAG	1 ovarian 2 breast 1 colorectal 1 unknown primary	35	No	Na	*BRCA2* positive
38	FS Del	635delA	1 ovarian 1 breast 2 colorectal 1 lymphoid 1 parotid 1 stomach	47	Breast cancer	36	Not tested
39	FS Del	0886delGT-ter220	1 ovarian 4 breast	66	Breast cancer	59	Not tested
40	FS Del	982del4	1 ovarian 4 breast	38	Malignant melanoma	38	Not tested
41	FS Del	982del4	1 ovarian 3 breast	39	No	Na	Not tested
42	FS Del	982del4	1 ovarian 2 breast 3 stomach 1 alimentary tract 1 lung/bronchial	42	No	Na	Not tested
43	FS Del	1493delA	1 ovarian 3 breast 3 prostate 1 unknown primary	39	No	Na	Not tested
44	Nonsense	1684C>T (Q486X)	1 ovarian 1 breast 2 alimentary tract 1 prostate	39	No	Na	Not tested
45	FS Del	2157delG-ter659	1 ovarian 3 breast 1 liver 1 pancreatic	41	No	Na	Not tested
46	FS Del	3773delTT-ter1182	1 ovarian 2 breast 1 uterus 1 colorectal 1 lung	49	No	Na	Not tested

47	FS Del	3908delTG-ter1231	3 ovarian	39	No	Na	Not tested
48	FS Del	3917delC	2 ovarian 1 breast 2 unknown primary	47 (Oophorectomy at 48)	No	Na	Not tested
49	FS Del	4075_4076delGT (STOP1284)	2 ovarian 1 breast 1 lung/bronchial 1 lymphoid 1 unknown	46	No	Na	Not tested
50	FS Del	4075_4076delGT (STOP1284)	1 ovarian 2 breast 1 bladder 1 cervical 1 unknown primary	60	Breast cancer	48	Not tested
51	FS Del	4626del5	3 ovarian 1 breast 1 prostate 1 lung/bronchial 1 oesophageal	52	No	Na	Not tested
52	FS Del	4626del5	1 unknown primary 2 lung/bronchial 1 skin tumour 1 leukaemia	38	Breast cancer	52	Not tested
53	FS Del	4638del4	1 ovarian 2 breast 1 lung/bronchial	41	No	Na	Not tested
54	FS Del	4705del4	3 ovarian 1 unknown primary 1 lung/bronchial	59	No	Na	Not tested
55	FS Del	5104delAA	2 ovarian 2 breast 1 endometrial 2 colorectal 1 stomach	51	No	Na	Not tested
56	FS Del	5104delAA	2 ovarian 1 breast 1 gastric 2 skin (basal cell)	39	No	Na	Not tested

57	FS Ins	5294_5295insA	2 ovarian 1 breast 1 colon 1 Hodgkin's lymphoma 1 bone marrow 1 leukaemia	54	No	Na	Not tested
58	FS Ins	5294_5295insA	2 ovarian 3 breast 1 colorectal 2 alimentary tract 1 stomach 1 oropharyngeal 1 pancreatic 1 bladder	42	No	Na	Not tested
59	FS Ins	5294_5295insA	1 ovarian 1 breast 2 colorectal 1 stomach 1 pancreatic	51	Breast cancer	49	BRCA2 positive

60	Nonsense	5507C>G-Ser1760ter	2 ovarian 1 breast 1 brain 1 lung 1 stomach 1 skin 1 liver	42	No	Na	Not tested
61	FS Del	5531delTT-ter1772	2 ovarian 1 breast 2 endometrial	42	No	Na	Not tested
62	FS Del	5531delTT-ter1772	2 ovarian 1 bladder 1 brain 1 unknown primary	43	No	Na	Not tested
63	FS Del	5578delAA (STOP1785)	1 ovarian 2 breast 1 prostate 1 unknown primary 1 oropharyngeal	44	No	Na	Not tested
64	FS Del	6174delT	1 ovarian 7 breast 1 prostate 1 bone and chrondroid tissue	50	No	Na	Not tested
65	FS Del	6174delT	2 ovarian 1 breast 1 prostate	39	No	Na	Not tested
66	FS Del	6174delT	2 ovarian 1 breast 1 colorectal 1 oropharyngeal	53	No	Na	Not tested
67	FS Del	6174delT	1 ovarian 4 breast	53	No	Na	Not tested

68	FS Del	6503delTT-ter2098	2 ovarian 3 breast	56	Breast cancer	51	*BRCA2* positive
69	FS Del	6503delTT-ter2098	2 ovarian 3 breast 1 endometrial 1 colorectal 1 prostate	52	Breast cancer	52	*BRCA2* positive
70	FS Del	6503delTT-ter2098	2 ovarian 2 breast	58	No	Na	Not tested
71	FS Del	6503delTT-ter2098	1 ovarian 3 breast 1 brain 2 oropharyngeal 1 stomach	36	No	Na	Not tested
72	FS Del	6503delTT-ter2098	1 ovarian 4 breast 1 alimentary tract 1 lung/bronchial	42	No	Na	Not tested
73	FS Del	6926delC	1 ovarian 1 breast 1 prostate 1 bladder	36	No	Na	Not tested
74	FS Del	7297delCT-ter2358	1 ovarian 1 breast 1 stomach 1 skin tumour	52	No	Na	*BRCA2* positive
75	FS Del	7297delCT-ter2358	1 ovarian 1 breast 1 uterus 1 stomach	52	No	Na	Not tested
76	FS Del	8803delC-ter2862	3 ovarian 1 unknown primary	39	No	Na	Not tested
77	Nonsense	9610C>T (R3128X)	3 ovarian 1 uterus 1 lung/bronchial	68	No	Na	Not tested
78	FS Del	2116delA	1 ovarian 2 breast 1 lung/bronchial 1 leukaemia	41	No	Na	Not tested

Table 4.13 Epidemiological data for samples positive for BRCA2 gene variants. *This table gives details on family history for 42 samples with protein-truncating variants in BRCA2 gene. 5 of these had prior breast cancer diagnoses; 3 are subsequently diagnosed with cancer (2 breast cancer, malignant melanoma) fairly soon after joining the study. The remaining 34 samples do not report cancer diagnoses at the time of writing (July 2013).*

Table 4.13 shows epidemiological data for 42 women positive for protein-truncating variants in *BRCA2.* These data show that 34 women with *BRCA2* variants are currently unaffected by breast or ovarian cancer and all of whom have a strong family history of various cancer types including breast and/or ovarian. The ages of women range from 35 years to 60 years at recruitment, therefore, representing pre-, peri- and post-menopausal women; of the women with prior or subsequent cancer diagnoses the age of diagnosis ranges from 36 years to 59 years.

Table 4.14 Epidemiological data for samples positive for gene variants in *BRIP1*, *PALB2*, *NBN*, *RAD51C*, *RAD51D* and *BARD1*

Sample	Variant type	Gene Variant	Epidemiological data				
			Family history	Age at Recruitment	Cancer Dx Y/N	Age Cancer DX	NHS Testing
79	FS Del	*BRIP1* c.2990_2993delCAAA	1 ovarian (71) 1 breast (41)	46	No	Na	Not tested
80	FS Del	*BRIP1* c.2255_2256delAA	2 ovarian (50, 74) 3 breast 2 lung/bronchial 1 oesophageal 1 brain	44	No	Na	Not tested
81	FS Del	*BRIP1* c.890delA	3 ovarian (68,68,72) 1 colorectal 1 bladder 1 gastric	51	No	Na	Not tested
82	FS Del	*BRIP1* c.128_131delTGTT	2 ovarian (49, 31)) 1 lung 1 colorectal	38	No	Na	Not tested
83	FS Del	*BRIP1* c.66C>A	1 ovarian (34) 6 breast (45, 57)	36	Breast cancer	39	Not tested
84	FS Del	*PALB2* c.2488delG	2 ovarian (55, 66) 3 breast (44, 44, 45) 1 lung/bronchial 1 stomach 1 leukaemia	46	No	Na	Not tested
85	FS Del	*PALB2* c.2167_2168delAT	2 ovarian (64,83) 1 breast (58) 1 cervical 1 unknown primary	64	No	Na	Not tested
86	FS Del	*PALB2* c.509_510delGA	2 ovarian (43, 89) 1 breast (55)	47	No	Na	Not tested
87	FS Del	*PALB2* c.172_175delTTGT	1 ovarian (62) 5 breast (47, 49, 83) 1 stomach 1 alimentary tract	38	Breast cancer	39	*BRCA1/2* test negative

88	FS Del	*NBN* c.1142delC	1 ovarian (56) 1 breast (68) 1 testicular 2 lung/bronchial 1 leukaemia 1 alimentary tract 1 oesophageal 1 oropharynx 1 unknown primary	62	Breast cancer	47	Not tested
89	FS Del	*NBN* c.657_661delACAAA	2 ovarian (40, 56) 4 breast (40, 48, 77) 1 stomach 1 lymphoid	39	No	Na	Not tested
90	FS Del	*RAD51C* c.158delC	2 ovarian (54, 70) 1 lung/bronchial	64	No	Na	Not tested
91	FS Del	*RAD51C* c.731delT	3 ovarian (51, 58, 76) 1 Leukaemia 1 Kidney 2 unknown primary	47	No	Na	Not tested
92	FS Del	*RAD51C* c.731delT	2 ovarian (87) 1 breast 1 colorectal 2 gastric 1 prostate 1 bone 1 brain 1 unknown primary	60	Breast cancer	57	Not tested
93	FS Del	*RAD51D* c.748delC	1 ovarian (54) 3 breast (63,68,70) 1 liver	46	No	Na	Not tested
94	FS Del	*BARD1* c.2291_2294delTAGA	1 ovarian (63) 1 breast (42) 1 brain 1 lung/bronchial 1 cervical	36	No	Na	Not tested

Table 4.14 Epidemiological data for samples positive for gene variants in BRIP1, PALB2, NBN, RAD51C, RAD51D and BARD1. *This table gives details on family history for 16 samples with protein-truncating variants in PALB2, NBN, RAD51C, RAD51D or BARD1 genes. 2 of these had prior breast cancer diagnoses; 2 are subsequently diagnosed with breast cancer fairly soon after joining the study. The remaining 34 samples do not report cancer diagnoses at the time of writing (July 2013).*

Table 4.14 shows epidemiological data for 16 women positive for a predicted protein-truncating variant in *PALB2*, *NBN*, *RAD51C*, *RAD51D* or *BARD1* genes. These data show that 12 of these women are currently unaffected by breast or ovarian cancer and all of whom have a strong family history of various cancer types including breast and/or ovarian. The ages of women range from 36 years to 64 years at recruitment, therefore, representing pre-, peri- and post-menopausal women. Of the women with prior or

subsequent cancer diagnoses the age of diagnosis ranges from 39 years to 57 years; Table 4.14 also includes the age at diagnosis for the affected relatives (where this is known). This shows that for *BRIP1* gene all those with gene variants show multiple ovarian and breast cancer cases in the family history and many of these are diagnosed under the age of 60 years. The family history includes both breast and ovarian cancer for every sample and in addition, many other cancer types are evident.

4.15 Clinical Relevance of Results

These results identify potential novel genes that could be targeted in follow-up studies. These data suggest that variants in *BRIP1*, *PALB2*, *NBN* and *BARD1* should be investigated in new studies. Whilst the frequencies of mutations in these genes are lower than *BRCA1* and *BRCA2*, three new genes *BRIP1*, *PALB2* and *NBN* show frequencies of 0.22%, 0.17% and 0.22% respectively and as these are in unaffected women with very strong family histories these gene variants suggest they are at a very high risk of developing ovarian or breast cancer. RAD51C and RAD51D have formerly been identified as breast and ovarian cancer susceptibility genes and these should continue to be characterised in large case-control studies. This study does not include a control group so population frequencies for variants in these genes cannot be examined, therefore positive predicted deleterious results must be interpreted with caution. Further studies are required, particularly because several of the volunteers are diagnosed with cancer very soon after recruitment in the study.

4.16 Example Pedigree

The example pedigree in Figure 4.13 demonstrates the transmission of gene variants from generation to generation.

The volunteer, indicated by the red arrow in the pedigree diagram (Figure 4.13), has not been tested in an NHS clinic. She has no living affected relatives and would therefore, not be eligible for NHS genetic testing. She has a strong family history of breast and ovarian cancer, in which 3 prior generations of women have had diagnoses of breast and or ovarian cancer under the age of 60. *BRCA1* gene mutations are inherited in an autosomal-dominant pattern, which is clearly evident in this diagram (Figure 4.13). This finding has clinical relevance in that the proband's risk of developing breast and or ovarian cancer before the age of 60 is extremely high. She is currently 50 years old with a family of 3 daughters. If she is tested clinically she could

be offered risk-reducing surgery (RRSO). In addition, she has 3 daughters, all of whom would have a high risk of testing positive for the same mutation. If the daughters are tested and found to be negative, their lifetime risk for developing breast and/or ovarian cancer would be calculated at levels closer to that of general population. If found to be positive they could be offered early screening until they have completed families after which they could be offered risk reducing surgery.

Figure 4.13 Pedigree diagram for sample No. 7 with NGS detected *BRCA1* frameshift variant c.4184_4187delTCAA

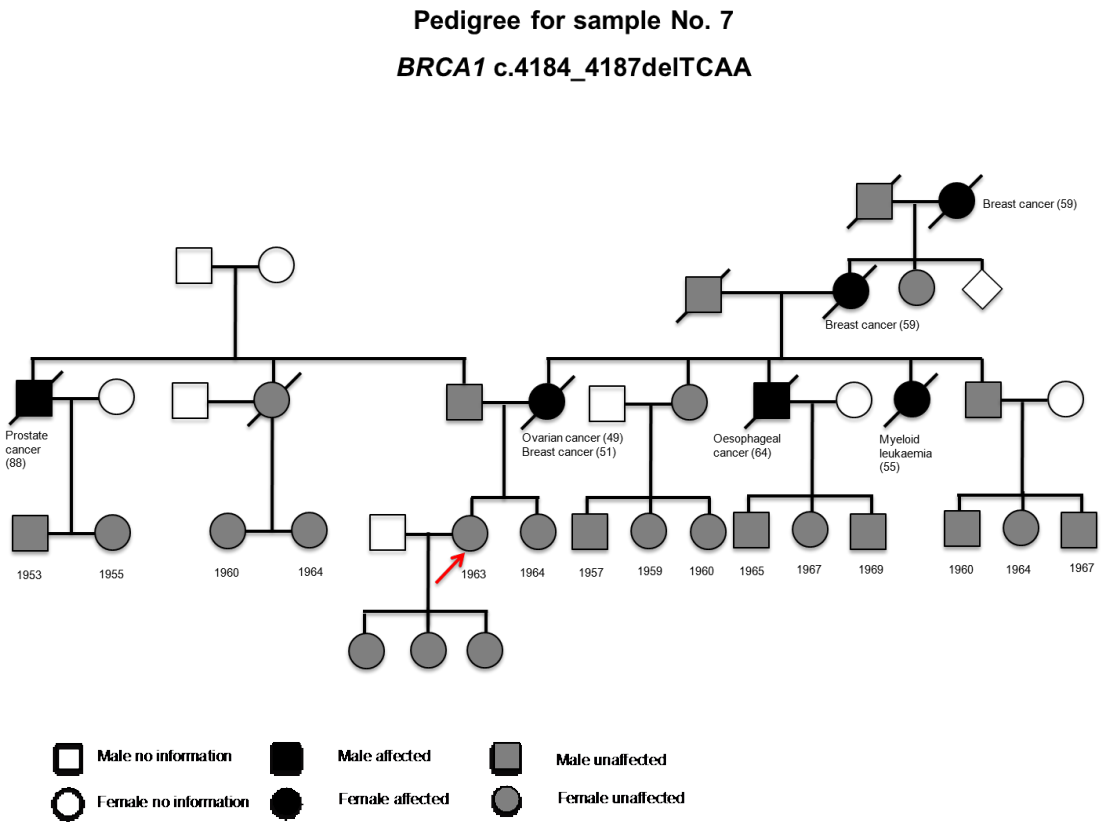


Figure 4.13 Pedigree diagram for sample No.7 with NGS detected BRCA1 frameshift variant c.4184_4187delTCAA. *This pedigree clearly shows BRCA1 transmission through the maternal side of the proband (indicated with a red arrow). No generations are skipped and there are approximately as many affected as unaffected women on the maternal side of the family; characteristics of highly penetrant gene variants inherited in an autosomal-dominant pattern. The proband's daughters have a 50% chance of inheriting the genotype from the mother.*

4.17 Discussion

This study continues with the same high-throughput targeted sequencing approach and employs this to characterise the frequency of 4 known and 5 novel ovarian cancer susceptibility genes in unaffected women from high-risk breast-ovarian cancer families. Multiplexed PCR on the Fluidigm Access Array platform is scaled-up further to allow for target enrichment and library preparation in all 9 genes in 2,300 subject samples.

4.18. Evaluation of the scaled-up targeted NGS approach used in this study

4.18.1 Target enrichment and library preparation

The target enrichment and library preparation are performed as 10 PCR reactions in each well of the Fluidigm Access Array. The methodology chapter describes which amplicons are mixed in each well. This is a complex, laborious and time-consuming part of the experiment, but is crucial in successfully sequencing all coding regions. Once this step is complete the rest of the target enrichment and library preparation steps are relatively straightforward and streamlined. In future experiments designing the primer mixing for highly multiplexed experiments could be improved by formulation of a computer program that could automate this process. A simple computer programming script could be written to follow the basic rules in mixing amplicons per well and that could also be applied to many different experiments.

4.18.2 Sequencing Quality Controls (QC) – sequence coverage

The importance of sequencing depth of coverage is discussed in the previous chapter. The minimum acceptable sequencing depth is previously established to be 30X and this level is used in filtering out reads. The excellent levels of coverage attained in this experiment allow for confidence in variant detection sensitivity and specificity.

4.18.3 The importance and pitfalls of NGS data filtering

Illumina sequencing uses sequencing by synthesis to determine the order of bases in the template DNA. Each DNA molecule is sequenced many times over and this sequencing depth allows for the accurate detection of single nucleotide variants and indels. This methodology therefore, produces vast quantities of data and can lead to sequencing artefacts unless read and variant filtering are performed.

If the first important step in NGS data analysis has already been identified as read alignment then the second most important step in variant annotation is filtering. The Genome Analysis Toolkit (GATK), described in previous chapters (McKenna et al

2010), includes tools for filtering variants, which will have the effect of reducing the vast number of variants detected in Illumina sequencing. Two of these tools from GATK, the Haplotype Score Test and the Read Pos Rank Sum test filtered out 11 of the 12 missing mutations and the low QD test filtered out the final missing mutation. Where haplotype score results are high these are filtered out as potential errors and sequencing artefacts. High scores suggest regions of poor alignment as the algorithm calculates segregating haplotypes at specific loci for each sample. If there are more than 2 haplotypes for a given sample then these are filtered out. Since the study population is concerned specifically with subjects with a high incidence of familial ovarian cancer these do not represent a random general population as would be the situation in a case-control study. Thus, perhaps the Haplotype Score test was not relevant here. The ReadPosRankSum test filters out variants that are consistently identified in the alternate allele at the end of reads. However, one major drawback here is that if the majority of reads are the alternate allele then this cannot be calculated accurately and good variant calls might be eliminated. This disadvantage is also relevant in some other variant filtering tools such as the MQ RankSum test, which compares read mapping qualities of the alternate allele and the reference allele. As only two of the filtering tools appeared to show up as reasons for eliminating the missing mutations it can be concluded that the filtering was mostly successful and only minor adjustments need to be made to improve the data analysis pipeline

4.18.4 Variant detection sensitivity

As the study participants are female relatives of ovarian cancer patients and part of a UK screening trial some are previously tested for mutations in *BRCA1* and *BRCA2*. This allows for further examination of NGS data to check for incorrect or missing mutation calls in these two genes. On inspection of clinical data 14 samples (out of a possible 73) are found to be discordant, i.e. those 14 samples were positive and were missed by NGS suggesting a poor variant detection sensitivity rate of 79.4%. Closer scrutiny in these 14 samples shows that these variants are in fact in the NGS data, but stringent filtering criteria removed the reads containing the variants.

The method of target enrichment used here will not detect large rearrangements such as deletions or duplications of whole exons. Thus for the *BRCA1*/*2* data these cannot be considered negatives neither can they be considered false negatives. These are the current limits of this particular methodology.

4.18.5 Variant detection specificity

As a validation method all predicted deleterious variants are sequenced using Sanger sequencing. 99 of the NGS variants out of a total 116 are validated (85.34%) in this way meaning that there are 17 false positives. N.B. only Sanger sequencing data from the 1st analysis is available at the time of writing. However, Sanger sequencing is still being conducted on the variants detected from the 2nd analysis with changed filtering parameters. In this study the filtering and QC levels are good, with 85.34% of positive samples being validated by Sanger sequencing.

In *BRCA1*/*2* genes there are several variants of uncertain significance (VUS). These VUSs may or may not be clinically relevant and in these data these VUSs have not been included in the positive results. In the clinical setting diagnostic laboratories detect around 5-10% of VUS (in *BRCA1/2*) even using Sanger sequencing. Clinics in these circumstances are still in debate around how these should be considered in terms of prediction of cancer risk. Thus, the samples containing VUSs sequenced in this study cannot be classified as true negatives or false negatives. Reassurance on the quality of these data can be taken from the accuracy estimates; as approximately 85% accuracy rate is in keeping with current levels in recently published data.

4.19 Evaluation of study design

One of the main advantages of a family-based study design is in the genetic enrichment for all types of risk allele, meaning that very rare high-risk and rare moderate risk alleles should be detected using this study design. This is in contrast to population-based case control studies that are unselected for family history and/or histological subtype that are more likely to detect the low risk common alleles. Additional advantages are in the circumvention of population stratification, a concept discussed in the previous chapter; and in the increase in statistical power and reduction in required sample sizes through using cases with a strong family history. Required sample sizes reduce with samples selected for family history as well as reducing with increasing allele frequency; this is especially marked where there are affected mother and sister in the family (Antoniou & Easton 2003).

4.20 Genetic variant prevalence and characteristics

For the purposes of this analysis deleterious mutations are classified as frameshift, nonsense and splice site alterations that are predicted to result in protein-truncation. This study did not examine the predicted missense variants. In this study the combined

prevalence of a deleterious mutation in the 8 genes containing identified variants is 5.26% (n=121) of which 4.39% (n=101) are *BRCA1* or *BRCA2* with the remaining 0.83% (n=19) attributable to the additional 6 genes. The ratio of *BRCA1* to *BRCA2* mutations is roughly equal in this study at 1.06:1 (*BRCA1*:*BRCA2*).

The prevalence figures in this study are strikingly low for the known breast/ovarian cancer susceptibility genes (*BRCA1*, *BRCA2*, *RAD51C* and *RAD51D*). This is a particularly surprising result since the sample set is genetically enriched with subjects from families containing multiple cancers, many of which include high numbers of ovarian and breast cancer cases. If ~50% of familial ovarian cancer is considered attributable to *BRCA1* and *BRCA2* genes, then these figure would be expected to be higher than 4.39%. After taking into account that the gene mutation is probably inherited in an autosomal dominant fashion, meaning that each family member has a 50% probability of inheriting the genotype, then this figure could be doubled to 8.78%.

This figure is still relatively low and several factors may explain this. For example, these families may be Lynch Syndrome families with mutations in DNA mismatch repair genes or it may be because there are still additional ovarian cancer susceptibility genes yet to be discovered. Another reason is that the cancer cases within these families may be sporadic cases that by chance happen to cluster within these families (Ramus & Gayther 2009).

Large genomic rearrangements such as large insertions or deletions of whole exons could not be identified in these data due to the choice of target enrichment and library preparation system. Mutations of this type could only be identified using genomic enrichment methods rather than one that creates tagged amplicons that only cover the coding sequence and splice sites of each gene. The missing large genomic rearrangements could in part explain the low prevalence figures for *BRCA1* and *BRCA2*. Ramus & Gayther (2009) analyse the prevalence of *BRCA1* and *BRCA2* mutations in 283 epithelial ovarian cancer cases selected with family history of ovarian cancer. They find that 37% of families have a mutation in *BRCA1* and 9% have a mutation in *BRCA2*. They perform MLPA analysis on a subset of the cases that have no detected coding sequence mutation; they identify 13 cases in the study with large genomic rearrangements. Ramus et al (2007) suggest that in UK ovarian cancer families around 18% are found to have large rearrangements in *BRCA1* gene. Another important factor to take into account here are the number of ovarian/breast cancer cases in each family. Ramus & Gayther (2009) report that the increasing number of cases in each family results in an increasing proportion of *BRCA1* and *BRCA2*

mutations. They report that in UK and US families with 2 ovarian cancer cases the prevalence of *BRCA1* and *BRCA2* mutations together is 27% rising to 63% in those where there are 3 or more ovarian cancer cases. Since this study analyses women that are disease free at the time of volunteering, then perhaps these are women with less than 3 ovarian cancer cases in the family, thus lower expected levels of prevalence in this study. In addition, Ramus & Gayther (2009) report that there are marked differences in prevalence figures in the UK and US populations. Some variation in prevalence figures would be expected between different studies, especially taking into account differences in ethnicity and sample sizes of individual studies. The most appropriate way to determine prevalence figures would be to combine data from many different studies conducted in different populations.

Age is relevant here in that as individuals' age and remain cancer free the probability that they have an inherited mutation diminishes. At a future date, it would be interesting to conduct analyses on these data examining the ages of individuals with detected mutations and the ages of individuals without mutations for those who develop cancer and those who remain cancer free.

In *BRIP1* Walsh et al (2011) detect 4 protein-truncating mutations in one peritoneal and 3 ovarian cancer cases. The ovarian cancer cases are 2 serous and 1 endometrioid histological subtypes. The mutations detected by Walsh et al (2011) are not the same as those detected here. In fact all *BRIP1* mutations detected in this study are not previously cited in literature. Certainly, this gene requires further investigation in very large case control studies and in family based studies in order to accurately estimate prevalence and penetrance levels and calculate odds ratios and relative risk. It is plausible that *BRIP1* may be found to be moderate penetrance in breast cancer but a moderate to high penetrance in ovarian cancer; these data suggest that *BRIP1* is potentially a novel ovarian cancer susceptibility gene.

PALB2 is the third most prevalent of the candidate genes in the study with 4 frameshift deletions detected in 0.17% of samples in the study. Two of these are previously reported in literature as connected with familial breast cancer c.509_510delGA (Casedei 2011) and c.172_175delTTGT (Hellebrand 2011). Two frameshift deletions, both in exon 5 are novel mutations found only in this study interestingly, one of these individuals has developed breast cancer and the other has a family history of breast cancer. Only larger case control studies will determine if these variants are specific to breast cancer only.

NBN gene represents the second most prevalent candidate gene in the study. This gene is not associated with ovarian cancer in previous studies and both of the mutations detected in this study are not previously reported in other studies as associated with either breast or ovarian cancer. These data suggest that this gene should be further investigated in larger follow-on case control studies.

The *BARD1* gene is already considered a candidate susceptibility gene due to its biological association with *BRCA1* (Ratajska et al 2013). Ratajska et al (2013) detect 3 probable deleterious variants in this gene in 109 non-BRCA1/2 breast and/or ovarian cancer patients. 1 predicted protein-truncating mutation, a frameshift deletion in exon 11 is detected in this study and this novel variant is not previously reported in literature; as such these data suggest this is a future candidate gene in follow-up studies.

In *RAD51B* no predicted protein truncating variants are detected and given that in the previous case-control study only 1 predicted deleterious variant is detected it is suggested that this gene is not a significant ovarian cancer susceptibility gene. In *RAD51C* and *RAD51D* fewer mutations are detected than in the previous study with only 3 mutations in *RAD51C* (0.13%) and 1 in *RAD51D* (0.04%). This study does not include a control group, which means that population frequencies cannot be assessed. These two genes were previously identified as breast and ovarian cancer susceptibility genes. The frequency of *RAD51C* is only lower by one variant compared to the other genes. Certainly, variants in all of the genes in this study are rare. In addition, this study could result in frequency estimations of half of that expected due to an autosomal dominant inheritance pattern. Or, the low frequencies could be due to the fact that these two genes are not predominant in familial ovarian cancer and may be more relevant to histological subtype and in the general population. Case-control studies include women with cancer diagnoses that do not have family history of the disease and this type of study is useful in identifying alleles in non-familial ovarian cancer. Case-control studies are especially useful when combined with epidemiological data including age of onset and histological subtype. This familial cohort study of unaffected women however, gives an indication of the likelihood of a woman harbouring gene mutations, within these candidates, with high familial risk.

4.21 Analysis of clinical relevance of study findings

Epidemiological data is gathered for 94 subject samples and includes, family history, age at recruitment, whether there is a cancer diagnosis, and if cancer is diagnosed at what age the diagnosis is made. In subjects with a *BRCA1* mutation 27 out of 36 study subjects are still unaffected at the time of follow-up. Of these 36, 5 have breast cancer

diagnoses at recruitment and 4 are diagnosed subsequently with cancer (2 breast, 1 ovarian and 1 cervical). The ages at recruitment range from 29 to 47 years suggesting that early detection strategies and risk prediction using genetic screening are likely to be of immense value in reducing mortality and morbidity in ovarian cancer. This is suggested as the women in this study are at a very high risk of developing ovarian cancer and the window of opportunity to identify these women will be during the pre- and peri-menopausal years. No subjects had a diagnosis of ovarian cancer at study recruitment. Similar statistics are noted for *BRCA2* in which the majority of women continue to be unaffected by cancer at follow-up. Again with the remaining variants in the additional 6 genes, where data is available 12 out 16 women continue to be unaffected at the time of follow-up.

Following this research, *BRIP1* is a potential novel ovarian cancer susceptibility gene and warrants further analysis. This gene could exhibit high-penetrance amongst mutation carriers. This discovery could have a major impact in the clinical diagnosis of prediction of risk of development of ovarian cancer. Furthermore, since these mutations were detected in those that were unaffected; then it would be plausible to extrapolate the prevalence found here and suggest that it could be doubled, this since this gene is likely to exhibit an autosomal dominant inheritance pattern.

Together these data suggest that three genes *BRIP1*, *PALB2* and *NBN* may potentially be novel ovarian cancer susceptibility genes and could be proposed for addition to clinical diagnostic screening for early detection and prediction of risk of ovarian cancer development. These could be included following further validation in large follow-up case control studies that examine the population frequencies of variants in these genes. In *RAD51C* 3 predicted pathogenic variants are detected supporting the significance of this gene in ovarian cancer. These data indicate that the contribution of genes in addition to *BRCA1* or *BRCA2* is highly relevant and the search for the remaining genetic contribution to ovarian cancer development is a valuable strategy. These data suggest that women with moderate to high-risk mutations are very likely to develop ovarian cancer and this is inferred from family based data; therefore, targeting these women in future follow-up studies for early screening would be highly beneficial in early detection of disease.

Large follow-up case-control studies will allow for the accurate prediction of the contribution of these novel ovarian cancer susceptibility alleles. Large numbers of samples, perhaps from several combined studies in population-based research, will result in the accurate calculation of odds ratios and relative risk. Only then can these

move to clinical genetic screening. Even then women will require counselling on the level of lifetime risk of developing disease. Discussion should centre on what risk is large enough to undergo RRSO.

Ethical discussion on the women that are at substantially increased risk should centre on the fact that this is in the research stage and that there may be specific and immediate implications in terms of the clinical responsibility towards those women in the UKFOCSS screening study. Ethical debate needs to urgently take place in order to be able to advise the women in this cohort that have been found to harbour genetic variants that are either currently known to increase risk of ovarian cancer development or could be found to increase ovarian cancer risk. This is to allow for women taking part in UKFOCSS to be able to make informed choices about their current and future health.

4.22 Conclusion

In conclusion, these data suggest that there are likely to be additional alleles for epithelial ovarian cancer and these data give insight into which further genes could be examined, for example, those genes in the FA Pathway do appear to be especially relevant as do those in the *BRCA1* pathway e.g. *ATM*, *ATR*, *MRE11* and those associated with *RAD50*. This reveals clues to further candidate genes. These data indicate that *BRIP1*, *PALB2* and *NBN* could potentially be included in clinical genetic screening, in the future, for early detection and improved risk prediction, which can only be introduced following further validation in large follow-up case-control studies.

Chapter Five

General Discussion of thesis

5.1 Next generation sequencing approaches for the identification of cancer susceptibility alleles

The introduction of next generation sequencing approaches is heralding a new era in genetic association studies for complex diseases, including epithelial ovarian cancer. This rapidly advancing technology is expanding the potential of genetic studies and enabling affordable high-throughput research allowing for the identification of rare, moderate and common cancer risk alleles.

5.1.1 Progress in technology during this thesis

Chapters 2, 3 and 4 carefully chart the progress in NGS approaches since the start of this thesis in 2009. Figure 5.1 is a flow diagram that depicts the progression of technology and the approaches used throughout the 3 years from 2009 to 2012.

Figure 5.1 The development of NGS approaches in this thesis (2009-2012)

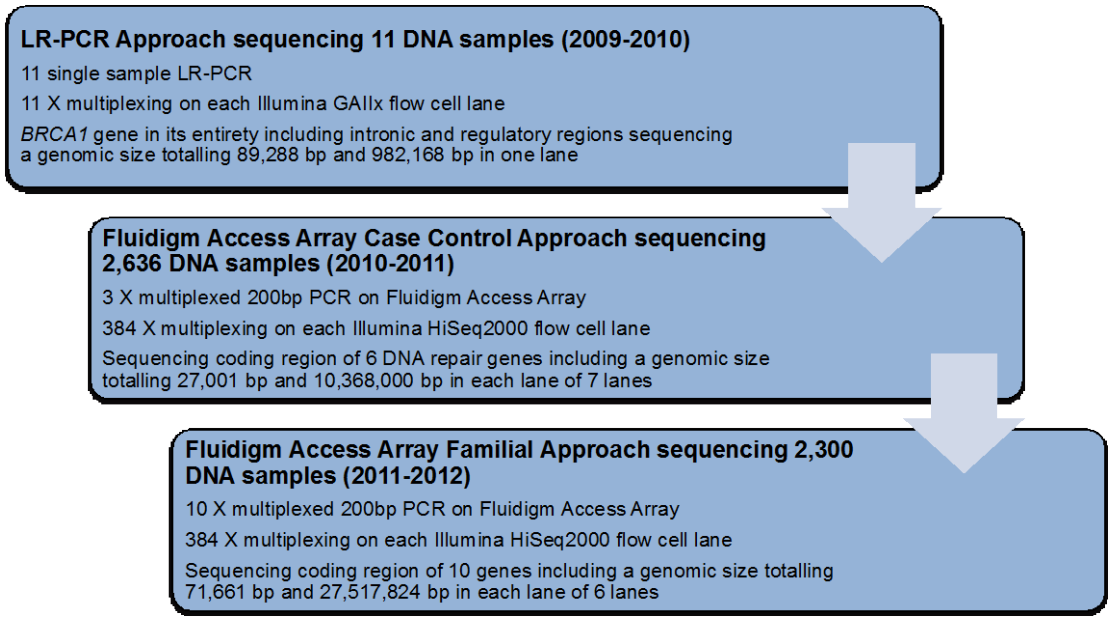


Figure 5.1 The development of NGS approaches in this thesis. This flow diagram charts the progress and development in the NGS approaches used through this thesis. This diagram shows how the increasing sequencing capacity of Illumina platforms is utilised to increase both throughput and genomic region investigated. LR-PCR = Long Range PCR, GAIIx = Genome Analyser IIx, bp=base pairs.

During this progression the time it takes to prepare libraries has been reducing along with the cost per base to perform target enrichment, library preparation and sequencing runs. Importantly, these improvements are not detrimental to depth of coverage levels due to the increasing capacity of NGS technology.

Not only do technological improvements make large genetic association studies more feasible in terms of time and cost per sample; they also introduce novel approaches in the detection of rare, moderate and common variants. Large DNA sequencing studies targeting candidate genes, the whole exome or whole genome are now possible enabling the identification of rare variants (with minor allele frequencies <1%), which have been almost impossible to find previously using the GWAS approach (Cirulli & Goldstein 2010).

5.2 Selecting the appropriate sequencing approach and study design in genetic studies

The approaches discussed here can be applied to a variety of different genetic diseases not only cancer. Most importantly, there are specific genetic characteristics that dictate the sequencing approach and study design most fit-for-purpose. For example, different sequencing approaches and/or different study designs may be required whether investigating for causal or predisposing variants in Mendelian diseases or complex diseases (Cirulli & Goldstein 2010). In addition, the mode of inheritance in the genetic disease may also be relevant in experimental design (Bamshad et al 2011).

5.2.1 Whole exome sequencing

Bamshad et al (2011) in a review paper describe both experimental design and analysis issues in using exome sequencing for the identification of genes in Mendelian diseases. Mendelian disease can be defined as those that are inherited in either a dominant or recessive pattern; most often these diseases are caused by variants in a single gene leading to a high risk of developing the condition. These same issues in design and analysis may also be relevant in complex diseases like cancer as alleles that predispose to cancer can be transmitted in a Mendelian fashion. For example, the pedigree outlined in Chapter 4 with the *BRCA1* mutation suggests that the mutation is inherited in an autosomal dominant pattern (refer to Figure 4.13). Bamshad et al (2011) suggest that exome sequencing is useful in determining both very rare (MAF <

0.1%) and rare alleles (MAF <1%) as this method should be able to detect all of these alleles in the sample. Tackling experimental design and data issues should also make exome sequencing more fruitful in dissecting these rare alleles in complex common disease and is thus not exclusively relevant to Mendelian or monogenic disease.

The whole exome can be defined as the full set of known coding exons in the genome, which consists of around 1% of the whole genome (Cirulli & Goldstein 2010). One of the predominant caveats in whole exome sequencing is the level of background noise. Background noise is derived from the abundant common polymorphic variants that have no effect on disease and the many sequencing artefacts found in NGS data. For example, Bamshad et al (2011) report that in European American DNA samples as many as 20,000 single nucleotide variants are detected by exome sequencing, the majority of which are previously known. They suggest that there are different approaches in identifying the relevant alleles within this context and consider differences in inheritance pattern, the population studied (i.e. family pedigrees or unselected cases and controls), how heterogeneous disease loci are and if the disease is caused by a transmitted variant or *de novo* variant. In addition, these issues are important considerations in defining appropriate sample sizes for satisfactory statistical power.

5.2.1.1 Filtering

Many of the non-pathogenic single nucleotide variants present in the human genome can be filtered out in several ways. Firstly, by using data publicly available on dbSNP and 1000 Genomes databases. Secondly, by filtering according to minor allele frequency (MAF); Bamshad et al (2011) suggest including only variants with a MAF of 0.1% or less in dominant inheritance disorders as these are likely to be very rare in the population. Subsequent to filtering, the remaining variants can be separated further by gene function, role or relationship in a biological pathway or by prioritising according to variant type; for example those that are predicted to be deleterious (frameshift, nonsense, splice site) may be more relevant than the predicted missense variants.

5.2.1.2 Inheritance pattern

Diseases that are inherited in a dominant or recessive pattern may require different levels of coverage. Intuitively, it may be expected that identifying homozygous alleles, using exome sequencing might be easier to detect and require lower sequencing coverage. For example, 40X depth may be sufficient (for research purposes) for the detection of homozygous variants sought will be homozygous. It follows then that for dominant inheritance diseases a higher level of coverage would be necessary; perhaps 80-100X since the variants sought will be heterozygous. However, in recessive disease, if variants were compound heterozygous the level of read depth required would be that of dominantly inherited disease. The reason for this is the proportion of reads for each allele. In heterozygous variants 50% of the reads should be seen for each allele, whereas for homozygous 100% of the reads should be the same. Sulonen et al (2011) in a study comparing exome capture techniques for exome sequencing suggest a coverage of 11X for 99% accuracy of calling heterozygous genotypes when these are compared to known SNPs from GWAS. However, their data show that less than half the region covered shows coverage levels of >10X; and this is the case across the board for all target capture platforms. As a rough estimate using exome sequencing for detection of novel SNPs a level of 20X coverage might expect to accurately detect ~95% of single nucleotide variants. At 4X then, less than 50% would be likely to be detected.

Conducting whole exome sequencing is still limited in terms of throughput and cost, so this type of experiment may not reveal the patients with the causative or predisposing mutations. This is where experimental design can assist further by using a genetically enriched sample set to increase the probability of detecting variants and reduce the sample sizes required to reach statistical power. Cirulli & Goldstein (2010) suggest two approaches in the identification of pathogenic alleles. (1) Exome sequencing of the affected members of families with several cases of the disease (2) 'Extreme trait sequencing', also known as targeted genotyping, refers to the study of subjects from the two extremes in distribution for a phenotypic trait. If using exome sequencing in ovarian cancer genetic association studies the most appropriate approach may require genetic enrichment by targeting those subjects with early onset disease in families with multiple cases (3 or more) of ovarian cancer only. Thus, combining a family based approach with targeted genotyping.

5.3 Current research progress in the discovery of genetic risk alleles in epithelial ovarian cancer and clinical relevance

In addition to *BRCA1* and *BRCA2* two new genes are identified as highly penetrant and predisposing to ovarian cancer, *RAD51C* (Meindl et al 2010) and *RAD51D* Loveday et al (2011). Rafnar et al (2011) detect a frameshift mutation in *BRIP1* that increases risk of ovarian cancer with an odds ratio of 8.13. However, these new genes are very rare and still only represent around 40% of inherited ovarian cancer risk. In this thesis 3 novel genes are identified as potential ovarian cancer predisposition genes (*BRIP1*, *PALB2* and *NBN*) and these genes require targeting in follow-on studies. This progress in identification of novel genes, suggests that continued searching for new alleles is worthwhile; whilst these are rare alleles, they could result in more than 10X increased lifetime risks of developing the disease (Gayther 2012). As such these genes may have important relevance in the clinical management of ovarian cancer cases and in risk prediction and early detection.

Pharoah et al (2002) suggest the genetic susceptibility to breast cancer is polygenic in nature and these multiple variants include rare (<1% MAF), uncommon (1% -5% MAF) and common (>5% MAF). Identifying these alleles through large sequencing studies will contribute to risk prediction estimates and in the development of mathematical algorithms.

Determining the functional role of these inherited risk alleles may be relevant in defining the biological mechanisms of ovarian cancer development and thus, these alleles are potentially novel targets for early disease detection. For the coding variants in this research, this is easier as it is possible to assess the effect of these changes on the function of the encoded proteins. In non-coding changes, such as SNPs found in GWAS studies, this will involve the use of further functional assays or fine mapping of genes. In addition, the determination and understanding of the interactions between low-risk alleles and environmental factors and other genetic factors is important in assessing cancer susceptibility (Gayther 2012). When the full panel of ovarian cancer susceptibility genes are determined it will be possible to accurately estimate an individual's specific ovarian cancer risk and offer the best most tailored and personalised medicine.

5.4 Personalised care in ovarian cancer: screening, early detection and targeted treatment

Personalised cancer care involves prevention, via early detection and screening of high-risk patients as well as specific treatment regimen tailored to particular histological subtypes. An in depth understanding of the molecular signatures and genetic pathways in ovarian cancer will certainly result in improved more targeted treatments for the disease similar to recent developments in breast cancer treatment.

Since genetic screening in *BRCA1*/*BRCA2* can steer decisions in both prevention and treatment of breast/ovarian cancer it is vital that this screening is the most comprehensive that technology allows. Using next generation sequencing approaches will undoubtedly give greater accuracy and greater genomic coverage. New technologies are faster and cheaper than those currently in use in clinical genetics laboratories, which will result in a larger proportion of women being offered genetic testing. When the treatment option is the radical bilateral salpingo-oophorectomy it is important to ensure against false positive results and also to be sure which mutations pose the most risk in order to provide the most accurate risk assessment.

This thesis has a significant impact on personalised medicine in several ways:

1. It identifies potential novel cancer susceptibility genes that could be incorporated into clinical genetic screening once validated.
2. It presents a viable screening approach in which high-risk women can be identified at a very early stage when prevention strategies can be implemented.
3. It has identified likely targets for therapeutic approaches. For example, those variants that are detected in genes with roles in homologous recombination may be sensitive to PARP inhibitors.
4. It has established and refined robust protocols for both high-throughput research projects for the prediction of disease risk related susceptibility alleles for any complex disease, not only cancer; as well as establishing a rapid and affordable system for mutation detection in the clinical setting. Thus, allowing for a much larger population of women to benefit from genetic screening for risk prediction and early detection of ovarian and/or breast cancer.

5.4.1 Personalised medicine

Personalised medicine is a branch of medical practice that focuses on the specific genetic profile of an individual (or an individual's disease) in order to guide the diagnosis and treatment of disease arising in that individual. In a paper produced by the European Science Foundation (ESF) Forward Look (2012) entitled 'Personalised medicine for the European citizen' at www.esf.org; the ESF suggests that personalised medicine is the provision of tailored healthcare specific to each person and that is conducted through prevention, diagnosis and treatment. However, they suggest that the term is often defined differently and additional terminology can be used to describe similar medical practice. There are distinct differences between the terms 'personalised medicine', 'genomic medicine' or 'stratified medicine'. Genomic medicine refers to sequenced genome data whereas the meaning of stratified medicine suggests specific populations of patients whom benefit from specific treatments. For example, breast cancer patients with *HER2* positive breast cancer respond to the drug Herceptin, whereas *HER2* negative do not. As a treatment regimen the term 'targeted treatment' is often used.

5.4.2 Predictive medicine

By contrast, predictive medicine focuses on assessing the probability that individuals may develop disease at some point in their lives. Often the result of this is an estimate of lifetime risk (i.e. the likelihood of disease developing by the age of 70 years). Predictive medicine includes genetic screening in high-risk groups and often involves DNA sequencing. In certain circumstances this can be conducted prenatally or on the neonate if early treatment strategies are important. For example for Familial Adenomatous Polyposis (FAP) preimplantation genetic diagnosis (PGD), prenatal diagnosis and neonatal or childhood DNA testing are available due to the autosomal dominant pattern of inheritance and almost complete penetrance (Douma et al 2010). Preimplantation genetic diagnostics (PGD) are conducted in very specific and rare circumstances; this is where embryos are screened for genetic diseases before implantation using in vitro fertilisation (IVF). Specific reasons for conducting PGD are usually due to severe monogenic diseases and can circumvent anxieties around decision-making on termination of pregnancy. PGD is now available for those with mutations in *BRCA* genes. This is performed on embryos of patients positive for *BRCA1* or *BRCA2* and leads to the selection of embryos that are tested negative for *BRCA* mutations (Wilkinson 2012).

PGD is also used in determining sex of embryos in the specific circumstances of X-linked genetic diseases in which female embryos are selected for implantation (Pray 2008). For example, the X-linked genetic disorder Duchene's Muscular Dystrophy (DMD) is an X-linked Mendelian recessive disease, which almost never occurs in females due to the likelihood that offspring will not inherit two copies of the recessive mutant allele responsible for the disease. In the UK all PGD is closely monitored and agreed by the Human Fertilisation and Embryology Authority (HFEA). The HFEA also regulate embryological research in the UK (www.hfea.gov.uk)

5.4.4 Advantages and disadvantages of personalised and predictive medicine

In practice predictive and personalised medicine often go hand in hand. Many of the obvious advantages of this contemporary approach to medicine include focusing on prevention strategies reducing not only mortality rates, but also reducing morbidity and ultimately reducing cost of treatment. This is found in the advice given to the population. For example, if an individual is assessed to have a genetically elevated risk of heart disease that individual can make informed lifestyle decisions to help minimise that risk. Diagnoses of single gene disorders, such as cystic fibrosis, are one of the successes of predictive medicine, in which early diagnosis results in improved and earlier treatment (Farrell et al 2008).

Since many diseases particularly complex ones like cancer are not caused solely by a gene mutation, predicting the likelihood of developing a specific cancer is inherently problematic. Other factors, including environmental and lifestyle issues play a major role in carcinogenesis.

Genetic counselling prior to taking the decision to have a predictive genetic test is required; this must include how to deal with the results, whether these are positive or negative. Patients need to be fully informed in the consequences relevant to the specific test or disease. Issues and questions that may need to be addressed in counselling might include the following:

1. The test may not reveal any useful information at all – as only around half of the genetic predisposition to ovarian cancer has been currently elucidated
2. If the test is positive, how will the patient deal with the possibility that the gene mutation has been (or could in the future) be inherited by offspring?

3. The current limitations of the technology employed to detect genetic mutations. It is shown here that there are false positives and false negatives. Patients need to be made aware of this.
4. There may be data protection issues and other ethical issues in terms of employment or insurance policies.
5. If found positive, how does the patient feel in terms of their responsibility to other relatives that may also harbour a faulty gene? Should they share this information with them or not? Is it their moral responsibility?
6. If found positive, what can the patient be offered for reducing the risk of developing ovarian cancer? Surgical options are radical personal choices. Is regular screening enough? And can regular early screening in fact increase anxiety by highlighting risk in daily life?

Other disadvantages include concerns in the commercial 'over-the-counter' availability of genetic tests for genetic diagnosis that may hamper progress in this area. These commercial tests threaten progress as they bypass health professional consultation. Current medical practice in genetic screening includes genetic counselling on the results, whether these are positive or negative. Reduction in the price and the commercial race amongst private companies suggest that these 'over-the-counter' tests could be in place to those that request it (Hawkins & Ho 2012).

5.5 The translation of NGS into clinical genetic screening

The translation of NGS into the clinical setting for genetic diseases such as rare monogenic disorders is far in advance of cancer and other complex genetic diseases. In this area of genetics up to half of the causative genes have already been discovered for some 7,000 rare diseases that are caused by a single-gene. In fact, Boycott et al (2013) in their recent review suggest that the remaining half will be discovered in less than 7 years.

By contrast, in multi-gene analysis for cancer predisposition, the pace of progress is slower, but steady due to the complex nature of cancer syndromes (Domcheck et al 2013). The progress of the use of NGS technologies into diagnostics is assisted by the introduction of new personalised sequencers. Illumina have introduced the new MiSeq sequencing system, which is a fast bench top personalised sequencer that can produce 15Gb data per run and 25 million sequencing reads in just 2 days. This fully automated system (from sample to analysed data) produces 300 bp paired-end reads

making its use in clinical diagnostics very attractive and as such the technology is ready for use in this setting. This introduces the potential of moving away from the current practice in clinical genetic testing, which involves testing one likely gene at time towards a multi-gene approach (Domcheck et al 2013). Indeed, there are already commercially available tests, which are either cancer or disease specific that include a panel of genes. These tests may be most useful in the better-defined areas of cancer predisposition, for example HPNCC, in which the causative genes are already discovered. In this instance NGS technology provides a rapid cost-effective genetic screening method. In most breast and ovarian cancer this is not so clear-cut. This is because there are still more genes to discover, the penetrance of these genes varies widely and because breast and ovarian cancer are highly heterogeneous. Thus, clarification of the risks posed by novel alleles discovered in in epithelial ovarian cancer through this thesis and additional follow-on studies are vital before we can translate these into the clinical setting. We are still in the research phase and the translation of this research into diagnostics, whilst may be the ultimate aim, is still some years away. In my view, the introduction of OvaNext (Ambry Genetics) is premature and could cause major anxieties in women found positive for these tests. This panel of genes includes 17 genes with estimated increased cancer risks of 2-5 fold. Many of those included are not yet fully defined in research, for example, *BARD1*, *NBN*, *PTEN*, *PALB2* are included and these are not yet fully investigated in large studies. In addition, many of these genes are implicated in other cancers and patients should be offered full pre-test counselling for these cancers as well.

5.5.1 Pre-test counselling for multi-gene cancer predisposition clinical screening

Specific pre-test counselling would be required if patients are going to undertake multi-gene testing for cancer predisposition, which should be tailored to each individual. The problem with pre-designed panels is that genes are included that are either not relevant or have far reaching implications. *TP53* is often included and is known to be involved in many different cancers and therefore, is not specific. Given the choice, would most patients decline this test? It is vital therefore, that patients are offered full counselling in a way that enables them to make informed choices about which of their own genes they want tested (Domcheck et al 2013). Specific genes should be able to be excluded, rather than offering a set panel of genes designed as one size fits all approach.

5.6 Genetic Testing and Society

5.6.1 Ethical, moral issues and legal issues

5.6.1.1 The lifetime level of cancer risk: at what level should we advise patients to take action?

Is the answer to this question something that can be advised on or is this purely a subjective individual opinion? Within the scientific community questions like these warrant debate as they give insight into how best to advise patients in the clinic. If it is possible to assess fairly accurately a person's level of risk then it is necessary to consider this in terms of how this risk is managed. In clear high risk cases such as mutations in *BRCA1* that can result in lifetime risks of up to 87% breast cancer and 60% ovarian cancer this may be easier. However, if a person's risk is estimated to be at 30%, 20% or 5% management options may be far less obvious. For many, a 5% level of risk may be simply cause for increased anxiety for which no medical intervention is either available or useful. Radical risk reducing surgery such as RRSO may be considered excessive by some women when their lifetime risk is 30% or less or this may be an attractive reassuring option for those with far lower risks. The opinions will be very individual to each woman; whether they have children or not, age and other commitments may all influence a woman's feelings on the appropriate level of risk for medical intervention to take place. On the other hand, many people may be interested in mapping out their entire genetic blueprint, although this is likely to be (in most cases) of very little medical benefit. This is particularly relevant in the context of the abundant non-pathogenic variants previously noted in this chapter.

If a healthy woman with a minimum of 2 FDR/SDR were considered to have a risk of 10% of harbouring a mutation in a moderate-high penetrance gene would this level of risk be high enough to advise on risk reducing surgery or early monitoring of disease? The answer is probably not; but offering genetic testing could be useful in reducing anxiety (if found negative) or can give an accurate and specific estimate of lifetime risk if positive. This would only be possible once the genetic susceptibility to ovarian cancer is fully defined and this could still be a long way into the future.

5.6.1.2 Public fear of genetics

All too often genetic research presented in the media is centred on gene technologies and ‘designer babies’. Scientists are banded as ‘playing god’; in reality genetic research has nothing to do with eugenics. Media publicity can hamper scientific and medical progress by instilling such irrational fears in the wider population. These fears can only be harmful to scientific and medical research. Better, more informed and more accurate publicity is required to redress this balance; to allow the public to educate themselves on real genetic research which aims to improve and impact upon public health not create a supreme genetic composition of the human race.

5.5.1.3 Insurance companies and genetic testing results

The concordat and moratorium on genetics and insurance is a policy and practice document agreed between the government and the Association of British Insurers (ABI). The insurance genetics moratorium means that those buying insurance policies, apart from life insurance in excess of £500,000, will not have to declare results of a genetic test taken to predict disease. In addition, insurance companies are not allowed to request that a person take a predictive genetic test before granting insurance. However, this does not prevent them from requesting information on family history of disease and information on any diagnostic tests taken within families and these could include diagnostic genetic tests. Thus, if a close relative has been diagnosed with a genetic mutation in one of the known cancer predisposition genes, this could affect the ability to get insurance cover or the cost of that cover. This moratorium revised in March 2013 will remain in place until 2017. This moratorium is going to be reviewed in 2014.

As increasingly more knowledge is gained on the genetic susceptibility to disease the issues affecting insurance cover could become more complex. In the future could insurance companies request a potential customer take certain predictive genetic tests? How well will government safeguard the consumer in the future? The answers to these questions are currently unknown and therefore, health policy to introduce genetic testing in the general population must carefully consider these ethical and legal issues.

5.6.1.4 Employment and workplace discrimination

Issues within employment or the workplace concerning genetic testing are included in data protection legislation. The Genetics Commission was an independent public body that was set up in 1999 to give advice to government concerning ethical, social and moral issues in genetic testing. However, in 2010 this quasi-autonomous non-governmental organisation (Quango) was closed through the government's mass review on all Quangos. Whether organisations such as these are necessary for the protection of those seeking predictive genetic testing will require debate. These issues may simply be best addressed in government policies on employment and discrimination not require a separate genetics commission; indeed an organisation such as this may monitor and delay scientific progress in this area. Perhaps the responsibility could lie with scientists becoming self-regulating. These and related issues will require debate within the scientific community to allow for continued progression in and for the population to benefit from predictive genetic testing.

5.7 Ethical and moral discussion on feedback of genetic testing results from this research

The members of PROMISE (2016) agree that genetic testing results should not be returned to study volunteers and there are many reasons why these results could not be returned. (1) The novel variants detected in this study are not yet validated in follow-on studies, meaning that the full clinical relevance of some of these variants is currently not known. (2) Accurate risk estimates cannot be calculated until many more studies are conducted and the results combined. (3) Some of the *BRCA1* and *BRCA2* variants detected are variants of uncertain significance and it is not known how much these will affect cancer risk. (4) The clinical interpretation of moderate penetrance alleles is uncertain i.e. these may show relative risk between 2 and 5 (Domcheck et al 2013). (5) The study volunteers did not receive counselling in relation to genetic testing (as the trial is centred on screening) and these novel genes are identified following the end of the UKFOCSS screening study.

However, whilst the ethics are clear there remains the moral issue in withholding important information on a few women in the study. For example, the proband in the pedigree detailed in chapter 4 (Figure 4.13) has a mutation in *BRCA1* gene, which means she has a very high risk of developing ovarian cancer. In addition, she has 3 daughters all of whom will have a 50% chance of inheriting the mutation. If she had a

living affected relative she would be eligible for genetic testing and subsequently her daughters as well. In a case as clear as this, the issue of our moral responsibility as researchers is raised. Should we return to this debate amongst PROMISE members? The same could hold for the case-control study in that *RAD51C* and *RAD51D* are cancer susceptibility genes that result in a significantly elevated cancer risk. One recommendation from this thesis is, that this is a question that needs to be addressed.

5.8 Impact of this research on the health of the female population

The findings in this thesis may in part contribute to novel medical strategies focused on risk prediction and early detection of ovarian cancer as well as all cancer. In the near future more women will be able to be offered genetic testing for ovarian cancer (or breast cancer) susceptibility alleles as the cost of testing becomes increasingly more affordable. This larger population of women would include those women with a strong family history, whether they have a living affected relative or not, and would include women with cancer diagnoses under 60 years. The latter will become particularly relevant if those tumours of patients with novel DNA repair genes are found to be sensitive to specific chemotherapeutic agents such as PARP inhibitors.

However, the full impact of these research studies may not become apparent for some years to come as more and similar follow-on studies are performed, published and the resulting data combined to allow us to fully define the genetic structure of ovarian cancer susceptibility.

5.9 Conclusions and Future Work

This thesis assists in guiding experimental design for future studies as well as giving an indication of which genes to prioritise. Exome sequencing in genetically enriched sample sets from families affected by multiple cases of ovarian cancer and/or early onset disease will be extremely useful. However, exome sequencing is still not cheap enough to be able to sequence the very large sample sizes to enable the identification of rare variants. Therefore, the candidate gene approach will still be a valid and valuable one as this is the only way the appropriate level of coverage alongside the sample size can be achieved. Likely candidates would continue to be those related to *BRCA1* and *BRCA2* in DNA repair and all genes in the homologous recombination repair pathway, the Fanconi anaemia genes and the *RAD51* related genes.

Together the two high-throughput studies conducted here and reported on in this thesis probably represent the largest genetic studies in ovarian cancer to date, sequencing almost 5,000 DNA samples in total. Studies of this magnitude are now possible with next generation sequencing approaches. In one single experiment it is conceivable that numerous genes can be investigated in sequencing studies of large sample sizes and this approach is fruitful in discovering rare susceptibility alleles for epithelial ovarian cancer.

Chapter Six

Materials and Methods

6.1 Methodology for Chapter Two

Laboratory work was conducted at UCL and at Source Bioscience Plc in Nottingham.

6.1.1 DNA Samples

DNA samples from ovarian cancer cases are selected with known mutations in *BRCA1*. Mutation identification was conducted blinded.

6.1.2 Target Enrichment – Long Range PCR.

6.1.2.1 Primer Design

11 primer pairs are designed for *BRCA1* using NCBI Primer-BLAST program

This program uses Primer3 to design primers and runs a BLAST search to ensure primers returned are only those specific to the input template. Primers are between 20 and 30 nucleotides in length and have a GC content of 40-60%.

Primers are generated using the GenBank sequences: L78833.1 Homos sapiens *BRCA1* (BRCA1) gene, complete cds; and NG_012772.1 Homo sapiens Breast Cancer 2, early onset (BRCA2).

6.1.2.2 Search for the best performing DNA polymerase for Long Range PCR

Several different commercially available DNA polymerases are tested to find the best performing product for Long Range PCR; these include several polymerases from Kapa Biosystems, Finnzymes, Fermentas Life Sciences and Invitrogen. Enzymes are first tested using pooled genomic female DNA sourced from Promega.

6.1.2.3 PCR Amplification

Fragments ranging between 5kb and 10kb are generated using the Invitrogen SequalPrep™ Long PCR Kit with dNTPs following manufacturer's protocols with an input DNA concentration of 25ng/20µl reaction volume. Primers are included at a concentration of 0.5 µM for each forward and reverse primer. Cycling Conditions are as follows: Initial Denature of 94^{o}C for 2 min followed by 10 cycles of 94^{o}C (denature) for 10 seconds, 65^{o}C (variable dependent on primer annealing temperature) for 30

seconds and 68°C (extension) for 1min per Kb. Then 25 cycles of 94°C (denature) for 10 seconds then 65°C (variable) for 30 seconds and 68°C for 10 minutes + 20 seconds per cycle (auto extension); then a final extension at 72°C for 5 minutes and hold at 15°C. Below is a table 6.1 detailing the size of the genomic region covered for each gene and the total size of PCR products including overlap.

Table 6.1 Size of genomic region covered for BRCA1 and total PCR products including overlap

Gene	Size of Genomic Region	Total Size of LR-PCR Products (includes overlap)
	86.965	88.695

Table 6.1. Size of genomic region covered for BRCA1 and total PCR products including overlap

6.1.2.4 Capillary Electrophoresis

Capillary sequencing is used to sequence a section of each LR-PCR fragment to verify fragments were the gene required.

6.1.2.5 LR-PCR product clean up

10 µl LR-PCR products per reaction are placed into wells of a 96 well plate. 4 µl EXOSAP-IT is placed into each well. The plate is incubated for 15 minutes at 37°C; and incubated again at 80°C for a further 15 minutes.

6.1.2.6 Sequencing reaction set up

3 µl of cleaned up LR-PCR product is added to the wells of a 96 well plate. 3 µl the forward *or* reverse primers that are used for LR-PCR amplification are added to each well. 4 µl of Big Dye is added to each well. A control well is used into which standard M13 primer is added. The plate is vortexed and centrifuged briefly.

6.1.2.7 PCR reaction

The 96 well plate is sealed and placed on a thermal cycler with the following protocol:

Initial denature:

Rapid thermal ramp (1^0/second) to 96^0C, 96^0C for 5 min.

25 cycles of:

Rapid thermal ramp (1^0/second) to 96^0C, 96^0C for 10 seconds

Rapid thermal ramp (1^0/second) to 50^0C, 50^0C for 5 seconds

Rapid thermal ramp (1^0/second) to 60^0C, 60^0C for 5min

Rapid thermal ramp (1^0/second) to 4^0C, hold at 4^0C for until purification

6.1.2.8 Sequencing reaction SEPHADEX®clean up

10 µl PRESEQ_CL is added to samples in relevant wells of the sequencing reaction plate and centrifuged for 10 seconds. The content of the sequencing reaction plate is added to the respective wells of the clean-up plate. A 96 well semi skirted plate and is barcoded and clean up plate placed on top then centrifuged at 910g for 5 minutes to elute DNA. 10 µl of dH_2O is added to each well in the sequencing reaction plate to increase the volume to 20 µl.

6.1.2.9 Load ABI 3730

The bar-coded plate is secured into one cassette and this cassette is loaded onto the input stacker before starting the ABI 3730 sequencer.

The output sequencing trace (chromatogram) is analysed using software from Applied Biosystems (Sequence Scanner v1.0).

6.1.3 Library Preparation

One library is prepared for each sample analysed. First 11 PCR products for each sample are purified using ZR-96 clean and concentrator (ZymoResearch) and quantified with the Quibit Fluorometer (Invitrogen); samples are then pooled in equimolar quantities. Each sample is fragmented by sonication using a Bioruptor and purified again using Qiaquick Columns and a vacuum manifold (Qiagen).

End repair is then performed on the fragmented DNA samples. T4 polymerase and klenow enzyme are used to remove the 3' overhang and fill in the 5' overhang and results in blunt ended fragments, ready for 3' adenylation. Samples are again purified on Qiaquick columns and a vacuum manifold. An A base is then added to fragments using klenow fragment (3'-5' exo minus); this allows for the fragments to be ligated to the adapters, which have a T overhang at their 3' end. Samples are then purified again using QiaQuick MinElute columns. Subsequently, adapters are ligated to fragments to allow for them to be hybridised to the solid surface of the flow cell. Samples are then purified again and at this point are stored at -20^{o}C.

Samples are next purified further and selected for size using an agarose gel. This ensures that any un-ligated adapters are removed or those that may have ligated to each other; additionally this selects the appropriate size of fragments that will become the templates for cluster generation. A 2% gel is prepared with 400ng/µl of Ethidium Bromide. The whole of each sample (30µl) is added to wells of the gel leaving one

lane empty between the ladder and samples and between each sample. The gel is run at 120 volts for 60 minutes. Gel slices of 2mm are cut at ~300bp using the ladder as a guide. Each gel slice is then purified using QiaQuick gel purification system. The resulting purified samples are then enriched using PCR to amplify the size-selected fragments. Multiplexing is achieved via the addition of a 6 base index sequence at this stage. Therefore, for a paired end read 3 primers are required: one for each read and one for the index. The individual index sequences are added to each sample so that 12 samples can be sequenced in parallel. Illumina provide 12 index primers and these can be used for each lane of the flow cell. Following addition of the index sequence and enrichment, the samples are then purified again using QiaQuick columns (Qiagen).

6.1.4 Library validation

Before cluster generation, the libraries require validation and quantification and this is performed using the Agilent Bioanalyzer 2100. This system performs an automated sizing and quantitation step, delivering the information in a digital format. DNA is analysed via on-chip gel electrophoresis using just 1µl of sample each time; and repeated in double or triplicate. The data output reveals the size of fragments in bp and the concentration of DNA in nmol/l.

6.1.5 Library normalisation and pooling

The average concentration for each sample is calculated from the double/triplicate repeats. Samples are then be pooled in appropriate quantities to ensure an equal concentration of DNA for each sample.

6.2 Sequencing

6.2.1 Cluster Generation

Reagents for cluster generation are prepared, including hybridisation buffer, wash buffer, amplification buffer, linearization mix, blocking buffer and sequencing primer mix, as per Illumina GAII protocols (refer to the document supplied by Illumina: *Single Read Sequencing User Guide GA2 1004831*). Reagents are then loaded into the cluster station and the flow cell is loaded. Following cluster generation, sequencing primers are hybridised and sequencing is commenced within 4 hours.

6.2.2 Sequencing-by-synthesis

76bp single read sequencing with the Genome Analyser IIx (GAIIx) involves preparing and installing reagents and starting the GAIIx and then checking the Quality Control Metrics before completing the sequencing run (full protocols are available in the document supplied by Illumina: *Single Read Sequencing User Guide GA2 1004831*).

6.3 Bioinformatics and Data Analysis

6.3.1 Basic Local Alignment Search Tool (BLAST)

DNA sequences from capillary electrophoresis are identified and confirmed as correctly amplified genes using the on line alignment program BLAST (Basic Local Alignment Search Tool). This program searches for regions of homology between sequences, by matching an input nucleotide sequence to those in sequence databases.

Bioinformatics for NGS sequencing data is performed by Bioinformaticians at Source BioScience Plc and involves first de-multiplexing samples to give 11 separate files. Then reads are aligned to the reference sequence and variants detected by software programs. Two software programs are used on these data.

6.3.2 CLC Genomics

CLC genomics workbench is a commercially available software solution for next generation sequencing. This software package performs read mapping and indel and SNP detection. A trial version of this software is initially used to run a first data analysis on the pilot study data.

6.3.3 CASAVA and SAMtools

A second analysis of the data is performed as the first analysis only detected 3 indels. The parameters are altered so that reads are included that had coverage of <30 X. This means that reads are included that are in low coverage regions.

6.3.4 Third analysis

A third analysis is conducted for Exon 2 in sample Pr_B1 to ascertain why the 11bp deletion (189del11) is not detected.

The FastQC program is first used to run basic quality control checks on this sample giving basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N

content, sequence length distribution, sequence duplication levels and overrepresented sequences.

The Burrows Wheeler Aligner (BWA) program is used to align just sample Pr_B1.

The SAM (Sequence Alignment Map) Format and SAMtools are used for storage and manipulation of sequencing reads.

Picardtools is used to further manipulate files in SAM format, removing duplicate and overrepresented sequences including the indexes used to identify individual samples. Local realignment is performed using Picard using the Genome Analysis Toolkit developed by the Broad Institute.

The Genome Analysis Toolkit (GATK) Broad Institute tools are used from GATK to perform data processing, variant calling and manipulation and variant QC. These include the GATK IndelRealigner, multiple sequence alignment tool and the Unified Genotyper.

The Integrative Genomics Viewer (IGV) program is used for data visualisation and is also from the Broad Institute

Following QC checks, reads from sample Pr_B1 are aligned with BWA, and then duplicate sequences removed with Picard tools. Then cleaned up sequences are re-aligned with GATK IndelRealigner and variant detection performed on exon 2 Pr_B1 using GATK and visualised with IGV.

6.4 Methodology for Chapter Three

Laboratory work was conducted myself at UCL, Great Ormond Street Molecular Genetics Laboratory and at Source Bioscience Plc. in Nottingham, UK. Maria Intermaggio and Andre Kim also conducted laboratory work at University of Southern California (USC), USA. Dr Ed Dicks, at Strangeways Research Laboratory in Cambridge, UK conducted the Bioinformatics.

6.4.1 DNA Samples

Total number of samples available to sequence in this study is 1557 cancer cases and 1131 controls and these are sourced from other case control studies and ovarian cancer registries as follows:

Gilda Radner Familial Ovarian Cancer Registry (GRFOCR)

Some of these were previously screened for *BRCA1*/2 and in addition, there were 20 samples that had not been screened for *BRCA1*/2 genes this allowed for estimation of false positive and false negative rates.

MALOVA study provides early onset cases and cases with a family history or with serous histology and a set of healthy controls

UKOPS (UK Ovarian Cancer Population Study) cases are samples with family history and serous histology and a large set of healthy controls.

UKFOCR (UK Familial Ovarian cancer Registry) samples are ovarian cancer cases with a strong family history including at least one first degree relative with ovarian cancer.

POL NCI ovarian case control study samples from Poland have at least one first degree relative with ovarian cancer.

JAC Polish ovarian cancer study samples are healthy controls.

AOCS (Australian Ovarian Cancer Study) cases are all serous histological subtype and include a set of age-matched healthy controls.

Chapter 3 includes a summary table of all samples available for sequencing in the study and where they are sourced

6.4.2 Target enrichment

6.4.2.1 Primer Design for amplification of target regions

Primers were designed by Fluidigm Access Array Design Team to cover the coding region of six candidate genes: *SLX4, RAD51B, RAD51C, RAD51D, XRCC2* and *XRCC3*. Figure 6.1 shows a summary of the Fluidigm Access Arrays; the individual design summaries are in Appendix IV.

.

Figure 6.1. Summary of the Fluidigm Access Array design

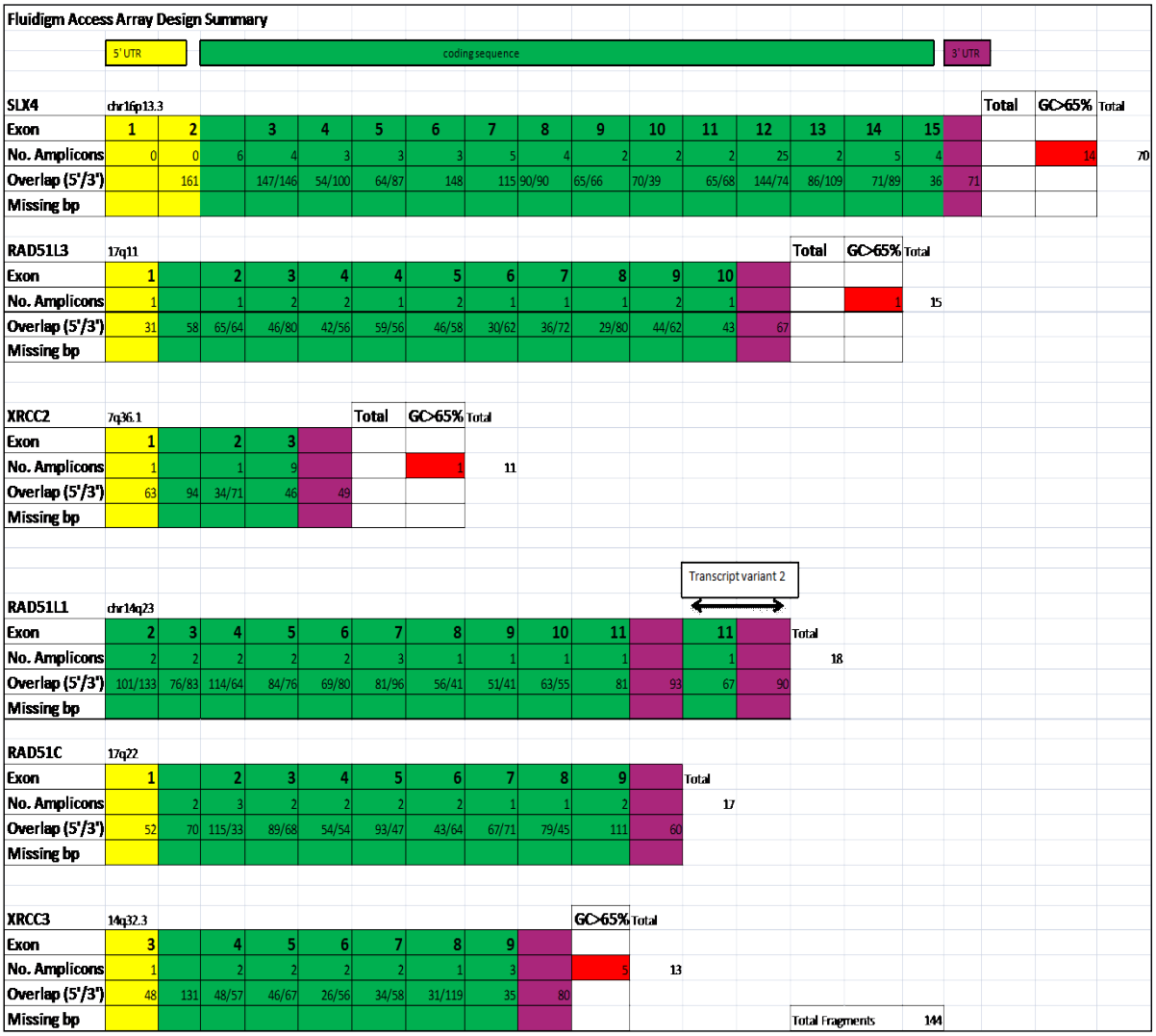


Figure 6.1. Summary of Fluidigm Access Array design. *This figure describes the number of amplicons prepared to include the coding region of each of the 6 genes. This figure also demonstrates the overlap in base pairs (bp) upstream and downstream to ensure that splice sites are also included. The figure also the number of amplicons (in red) in each gene with a GC% content >65%; this is flagged up as these amplicons may not amplify as efficiently as those with GC content <65%.*

6.4.2.2 Wet validation

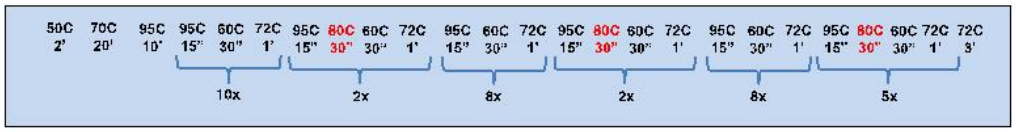
Fluidigm, using the protocol (Figure 6.2), conduct wet validation of the primer pairs:

Figure 6.2 Fluidigm wet validation protocol

- PCR Reaction Composition (total volume 5 µl)

Component	Stock concentration	Final concentration	per well
FastStart High Fidelity Reaction Buffer (w/ 18 mM $MgCl_2$)	10 X	1 X	0.50 µL
$MgCl_2$	25	2.70 mM	0.54 µL
DMSO	100 %	5 %	0.25 µL
PCR Grade Nucleotide Mix	10 mM	0.20 mM	0.10 µL
FastStart High Fidelity Enzyme Blend	5 U/µL	0.05 U/µL	0.05 µL
Access Array loading reagent	20 X	1 X	0.25 µL
BC primer pair	4.0 uM	0.40 uM	0.50 µL
DNA	60.0 ng/µL	10.0 ng/µL	0.83 µL
dH2O			0.98 µL
Inner primer pair	250 nM	50.0 nM	1.0 µL

- C0t Protocol



- Caliper analysis- 3 µl of sample + 12 µl of water (1:5 dilution)

Figure 6.2 Fluidigm wet validation protocol. *This figure shows the PCR reaction and cycling conditions for the wet validation protocol.*

6.4.2.3 Pooling of Primer Pairs

Primer pairs are pooled at 3plex (i.e. 3 PCR reactions per well) and prepared so that half of a 96 well plate (48) is used. Rows A-H are filled to column 6 with each row in the diagrams representing 1 well, as described in Figure 6.3.

Figure 6.3 Pooling of primer pairs

Column 1

A	RAD51C_t1_1	RAD51L3_t10_1	SLX4_t12_10
B	RAD51C_t1_2	RAD51L1_t8_1	SLX4_13r_3
C	RAD51C_t2_1	RAD51L1_t9_1	XRCC2_t3_6r_1
D	RAD51C_t2_3	RAD51L1_t2_2	SLX4_t2_37
E	RAD51C_t2_4	XRCC3_t1_1	RAD51L1_t6r_2
F	RAD51C_t3_1	RAD51L1_t5_2	SLX4_t5_3
G	RAD51C_t3_3	RAD51L1_t6_1	SLX4_t12s_2
H	RAD51C_t4_1	RAD51L1_t6_3	SLX4_t12s_5

Column 2

A	RAD51C_t4_2	RAD51L1_t1_2	XRCC2_t3_1r_1
B	RAD51C_t5_1	RAD51L1_t10_1	SLX4_t12s_4
C	RAD51C_t5_2	RAD51L1_t1_1	SLX4_t3_4
D	RAD51C_t6_1	RAD51L1_t11_1	SLX4_t2_34
E	RAD51C_t6_2	XRCC2_t2_1	RAD51L1_t4_1
F	RAD51C_t7_1	RAD51L1_t2_1	XRCC2_t3_8
G	RAD51C_t8_1	RAD51L1_t4_2	XRCC2_t3_1
H	RAD51C_t9_1	RAD51L1_t5_1	XRCC2_t3s_1

Column 3

A	RAD51C_t9_2	SLX4_t12_20	SLX4_t2r_2
B	RAD51L1_t3_1	XRCC2_t3_9	SLX4_t12s_6
C	RAD51L1_t3_2	XRCC2_t3_6	SLX4_t12_23r_1
D	RAD51L1_t7_1	XRCC3_t2_1	SLX4_t2_33
E	RAD51L3_t1_1	SLX4_t12_21	SLX4_t3_3
F	RAD51L3_t10_2	SLX4_t4_1	SLX4_t11_2
G	RAD51L3_t11_1	SLX4_t4_3	SLX4_t12_1
H	RAD51L3_t2_1	SLX4_t4r_2	SLX4_t11_1

Column 4

A	RAD51L3_t3_1	SLX4_t12_10r_2	XRCC3_t4_2
B	RAD51L3_t3_2	SLX4_t12_23r_2	SLX4_t5_2
C	RAD51L3_t4_10	SLX4_t12_15	SLX4_t7_8
D	RAD51L3_t5_1	XRCC2_t3_1r_3	SLX4_t12_14
E	RAD51L3_t6_1	SLX4_t12_16	SLX4_t5_1
F	RAD51L3_t6_2	SLX4_t12_17	SLX4_t7_3
G	RAD51L3_t7_1	SLX4_t12_17r_1	SLX4_t7_1
H	RAD51L3_t8_1	SLX4_t12_17r_2	SLX4_t8_4

Column 5

A	RAD51L3_t9_1	SLX4_t7_2	SLX4_t12_2
B	SLX4_t10_1	XRCC3_t4_1	SLX4_t14_3
C	SLX4_t10_3	SLX4_t13_4	XRCC3_t5_2
D	SLX4_t12_13	XRCC3_t2_2	XRCC2_t3_4
E	SLX4_t12_22	SLX4_t7_4	SLX4_t2_29
F	SLX4_t12_23	RAD51L3_T4r_6	SLX4_t2_35
G	SLX4_t12_23r_3	SLX4_t3_2	SLX4_t13_1
H	SLX4_t12_9	SLX4_t3_1	SLX4_t7_5

Column 6

A	SLX4_t12s_1	SLX4_t14_4	SLX4_t7_7
B	SLX4_t12s_3	XRCC2_t1_1	XRCC3_t6_1
C	SLX4_t13_2	SLX4_t8_1	XRCC3_t3_1
D	SLX4_t13_3	SLX4_t8r_2	XRCC3_t3_2
E	SLX4_t13_5	SLX4_t8_3	XRCC3_t7r_2
F	SLX4_t13s_1	SLX4_t7_6	XRCC3_t5_1
G	SLX4_t14_1	SLX4_t9_1	XRCC3_t7_1
H	SLX4_t14_2	SLX4_t9_2	XRCC3_t7_3

***Figure 6.3 Pooling of primer pairs**. Each image represents all rows in 1 column of the plate. Each row represents 1 well of a 96 well plate. Pools are created to include columns 1-6 of a 96 well plate*

6.4.3 Target Enrichment and Library Preparation

6.4.3.1 Overview of Multiplex Amplicon Tagging for Illumina on the 48.48 Access Array IFC

Sequencer specific tagged amplicons are generated in two PCR reactions. Each reaction is multiplexed at 3 pairs of primers in each well. All 144 amplicons are created on one Access Array Integrated Fluidic Circuit (IFC). Following the first PCR reaction, the products are harvested and used as a template in the second PCR reaction, during which the 384 barcode sequences are added along with the Illumina sequencing adaptors. Each IFC produces 144 amplicons in each of 48 samples each one uniquely indexed to allow for the pooling of all 6,912 PCR products.

Figure 6.4 Overview of Access Array System

Figure 6.4 Overview of Access Array System.

In summary, the Fluidigm Access Array is prepared as follows:

- Prepare sample pre-mix and sample mix
- Prime IFC
- Load IFC and run Load Mix Script in IFC controller AX – pre-PCR (60 min)
- Load IFC onto Biomark and run PCR (2.5 hours)

- Load harvesting reagents and run Harvest Script in IFC controller AX – post-PCR (1 hour)
- Transfer samples to a 96 well plate

The sequence tags and sample barcodes are attached as follows:

- Prepare sample pre-mix solution
- Prepare 100 fold dilutions of harvested PCR products to be used as a template
- Add sample pre-mix, 384 barcode and 1ul diluted harvested PCR products to 48 wells of 96 well plate
- Place PCR plate in FC1 thermal cycler ~ 40 min
- Store at -20^{0}C

6.4.3.2 Preparation of 20X primer solutions

20X primer solutions for 144 primer pairs are prepared. The table 6.2 below shows the primer dilution for one well (of 48 wells) with 3 primer pairs.

Table 6.2 Preparation of 20X primer solutions

Component	Volume (µl)	Final Concentration
CS1-TS-F (50uM)	2.0 µl per primer	1uM
CS2-TS-R (50uM)	2.0 µl per primer	1uM
20X Access Array loading reagent	5.0 µl	1X
DNA Suspension Buffer (TE)	83.0 µl	
Total	100.0 µl	

Table 6.2 Preparation of 20X primer solutions.

In a 96 well plate 4µl of each pooled forward and reverse primer pair are added at a mixture of 3 primer pairs in each well. This produces enough for 20 IFC arrays. The 20X primer solutions are divided into 5µl aliquots. This means that each 5µl aliquot is ready prepared for chip set up and to remove continued freeze/thaw cycles. Only 4µl is loaded onto the access array, but 5µl prepared to allow for dead volume

6.4.3.3 Prime the 48.48 access array

The Access Array is primed by injecting 300µl control line fluid into both of the 48.48 access array Integrated Fluidics Circuit (IFC) accumulators, named Contaminant Accumulator and Interface Accumulator in Figure 6.5.

Figure 6.5 The 48.48 Access Array Integrated Fluidic Circuit (IFC)

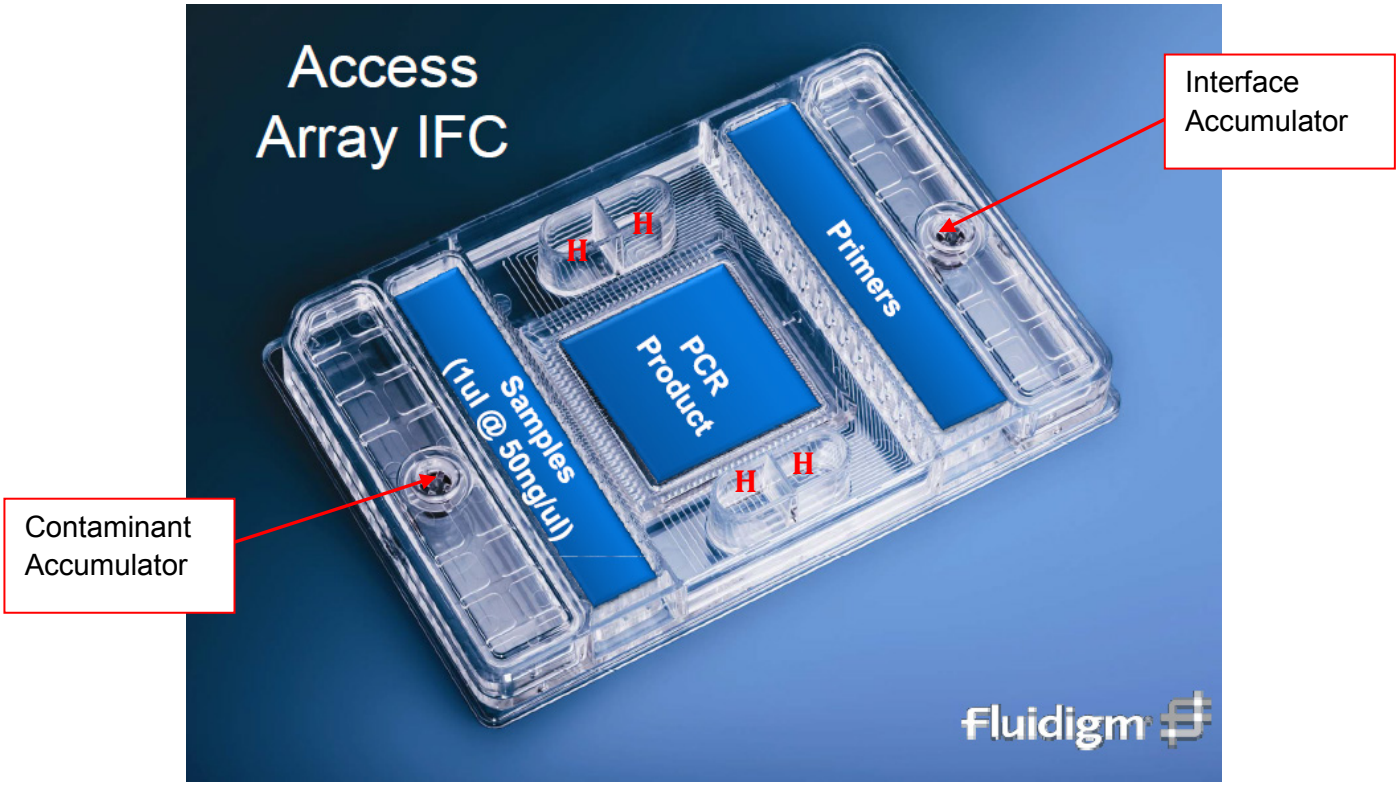


***Figure 6.5. The 48.48 Access Array IFC Chip**. This image shows the IFC chip platform. H = harvest reagent wells H1 to H4. Red arrows show were control line fluid is added to each accumulator.*

Then 500µl 1X Access Array Harvest reagent is added to the H1-H4 wells on the IFC and the Access Array is placed into the Pre-PCR IFC controller AX. Then Prime (151x) script is run on the IFC controller.

6.4.3.4 Preparation of Sample Pre-mix Solution

A master mix is prepared to a total volume sufficient for 60 reactions. The following components (Table 6.3) are combined to make the master mix.

Table 6.3 Preparation of sample pre-mix solution

Component	Volume (µl)	Volume for 60 Reactions (µl)	Final Concentration
10X FastStart High Fidelity Reaction Buffer without $MgCl_2$ (Roche)	0.5	30.0	1X
25mM $MgCl_2$ (Roche)	0.9	54.0	4.5mM
DMSO (Roche)	0.25	15.0	5%
10 mM PCR Grade Nucleotide Mix (Roche)	0.1	6.0	200 µM
5U/ul FastStart High Fidelity Enzyme Blend (Roche)	0.05	3.0	0.05 U/µL
20X Access Array Loading Reagent (Fluidigm)	0.25	15.0	1X
PCR Grade Water	1.95	117.0	
Total	4.0	240.0	

Table 6.3 Preparation of sample pre-mix solution.

The sample pre- mix solution is vortexed for 20 seconds and centrifuged for 30 seconds

6.4.3.5 Preparation of Sample Mix Solution

The sample mix solution is prepared by combining the following components (Table 6.4) in a 96 well plate.

Table 6.4 Preparation of sample mix solution

Component	Volume (µl)
Sample Pre-Mix	4.0
Genomic DNA (100ng/µL)	1.0
Total	5.0

Table 6.4 Preparation of sample mix solution

The sample mix solution is vortexed for 20 seconds and then centrifuged for 30 seconds.

6.4.3.6 Loading the IFC

Samples and primers are pipetted to the wells of the IFC chip. 4µl of sample mix is pipetted to each of the 48 sample inlets on the left of the chip and 4µl of 20X primer mix solution is added to each of the 48 assay inlets on the right side of the chip. The chip is placed into the Fluidigm Biomark and the C0t PCR protocol is run (Figure 6.6)

Figure 6.6 C0t PCR protocol

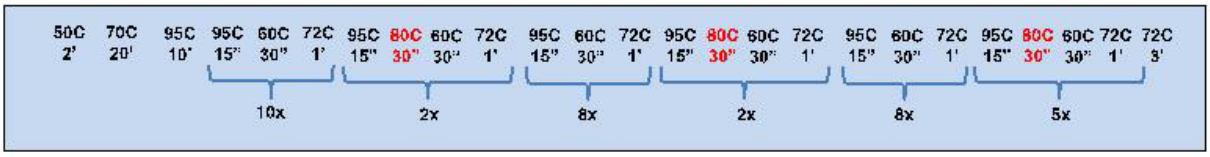


Figure 6.6 C0t PCR protocol.

6.4.3.7 Harvesting PCR products from the 48.48 Access Array IFC

In the post PCR lab 2 µl of 1X harvest reagent is pipetted into each of the sample inlets on the Access Array IFC. The remaining harvest reagent is removed from the H1-H4 wells. 600 µl of fresh 1X access array harvest reagent is added to H1-H4 wells. The access array is placed into the post PCR IFC controller and the script Harvest (151x) is run. When the harvest script is complete, the PCR products are removed and transferred from the sample inlets to the first 6 columns of a 96 well PCR plate.

6.4.3.8 Attaching sequence tags and sample barcodes

First, the sample pre-mix solution is prepared by combining the following components (Table 6.5) to produce enough sample pre-mix for 60 reactions. Then the sample pre-mix is vortexed for a minimum of 20 seconds and centrifuged for a minimum of 30 seconds.

Table 6.5 Preparation of sample pre-mix (attaching sequence tags and barcodes)

Component	Volume (ul)	Volume for 60 Reactions (ul)	Final Concentration
10X FastStart High Fidelity Reaction Buffer without $MgCl_2$ (Roche)	2.0	120.0	1X
25mM $MgCl_2$ (Roche)	3.6	216.0	4.5mM
DMSO (Roche)	1.0	60.0	5%
10 mM PCR Grade Nucleotide Mix (Roche)	0.4	24.0	200 nM each
5U/ul FastStart High Fidelity Enzyme Blend (Roche)	0.2	12.0	0.05 U/µL
PCR Grade Water	7.8	468.0	
Total	15.0	900.0	

Table 6.5 Preparation of sample pre-mix.

A 100-fold dilution of the harvested PCR products is then used as a template in the second PCR reaction. In a 96 well plate 99 µl of PCR water is added to 48 wells. Then 1 µl of harvested PCR product is added to each well. The PCR product dilutions are vortexed for a minimum of 20 seconds and centrifuged for a minimum of 30 seconds.

6.4.3.9 Prepare sample mix solutions

The following components (Table 6.6) are combined in a 96 well plate to prepare 48 individual sample mix solutions.

Table 6.6 Preparation of sample mix solutions (attaching sequence tags and barcodes)

Component	Volume (µL)
Sample Pre-Mix	15.0
Bidirectional 384 Barcode, Plate A1 (Fluidigm, PN 100-3772)	4.0
Diluted Harvested PCR Product Pool	1.0
Total	20.0

Table 6.6 Preparation of sample mix solutions

6.4.3.8 Thermal cycling

Thermal cycling is conducted in a 96 well plate using the following protocol (Table 6.7)

Table 6.7 Thermal cycling conditions to add sequence tags and sample barcodes

PCR Stages	Number of Cycles
95^{0}C 10min	1
95^{0}C 15 s **60^{0}C 30 s** **72^{0}C 1 min**	15
7272^{0}C 3 min	1

Table 6.7 Thermal cycling conditions to add sequence tags and sample barcodes.

6.4.4 Checking the barcoded PCR Products

PCR products are checked using an Agilent Bioanalyzer 2100. 1 µl of each reaction is checked using a DNA 1000 chip from Agilent following the manufacturer's instructions. The remaining products were stored at -20^{0}C

6.4.5 Pooling the PCR products for each IFC array

A harvest sample pool is created from each IFC array in order to prepare for purification and quantification. 8 pools are created, that is one pool for each IFC, by pooling 2 µl of each sample up to a total of 48 samples.

6.4.6 Purification of the pools

Purification of each individual pool is performed using AMPure XP beads. AMPure XP beads are removed from the fridge and allowed to warm to room temperature for 30 minutes. A fresh 70% ethanol solution is prepared by adding 3 mL of PCR water and 7 mL of 100% ethanol to a 15 mL tube. The tube of ethanol is vortexed for 5 seconds. The beads are vortexed for 10 seconds before adding the harvested pool and AMPure XP beads to one well of a 96 well plate (maximum capacity 300 µl) as detailed in Table 6.8

Table 6.8 Volumes of harvest sample pool and AMPure XP beads

Component	Volume
Harvest Sample Pool	96 µl
AMPure XP beads	180 µl

Table 6.8 Volumes of harvest sample pool and AMPure XP beads. *This table shows the proportions of harvest sample pool and AMPure XP beads used to purify the pools.*

The components, listed in Table 6.8, are mixed in the well by gentle pipetting and then vortexed for 10 minutes. The 96 well plate is then placed on the magnetic separator and allowed to sit for 3 to 5 minutes until the supernatant is clear and the beads are at the bottom of the well. The supernatant is removed using a pipette without disturbing the beads. 290µL of 70% ethanol is added to the well and placed on the magnetic separator again and allowed to sit for 3 minutes. The supernatant is removed without disturbing the beads, then a further 290µl 70% ethanol is added to each well and the plate left on the magnet and allowed to sit for 3 minutes. The supernatant is removed without disturbing the beads and the plate centrifuged briefly, then the resulting supernatant removed carefully. The plate is left to sit on the magnetic separator and allowed to air dry for a brief period of up to one minute. The beads are re-suspended in 40 µl of nuclease free water and then vortexed for 2 minutes. The plate is then placed on the magnetic separator and allowed to sit for 1 minute. The supernatant is removed and added to a fresh eppendorf tube ready for quantitation.

6.4.7 Quantitation and normalisation of pools

6.4.7.1 Quantitation

Each of the pools created from individual Access Arrays are quantified using the Agilent Bioanalyzer 2100.

6.4.7.2 Normalisation

In sets of 8 these pools are pooled further in equimolar quantities to form a final pool for one lane of the Illumina Flow Cell. Then each final pool (one lane) is quantified again using the Agilent Bioanlyzer 2100, in triplicate and the mean across these taken as the undiluted pool concentration. The required final concentration for sequencing libraries on Illumina HiSeq2000 is 10nM. Pools are diluted to this concentration ready for sequencing.

6.4.8 Dilution of pools to 10nM

Pools are diluted by adding an appropriate volume of water to a volume of DNA to achieve the desired concentration of 10nM, using the following formula:

Dilution factor = Initial Molarity ÷ Final Molarity (10nM)

DNA volume = Final Volume (50µl) ÷ dilution factor

6.4.9 Final quality control check

A final QC check is performed and each diluted pool is quantified in triplicate using the Agilent Bioanalyzer 2100. The actual final concentration is given as the mean across these triplicates.

6.5 Sequencing

6.5.1 Sequencing prepared tagged amplicons on the Illumina HiSeq2000.

The following diagram (Figure 6.7) describes how paired end sequencing of Fluidigm prepared libraries is performed on the Illumina HiSeq2000 platform. The amplicon produced on the Access Array contains Illumina paired end adaptors, Fluidigm custom sequencing primers primers and individual barcode sequences. The paired end adaptors (PE1 and PE2) are required for the amplicon to hybridise to the flow cell surface in each of the forward and reverse directions during the sequencing procedure. Read 1 is the forward read (PE1 and CS1), read 2 is the index read (BC and CS2rc) and read 3 is the reverse read (PE2 and CS2).

Figure 6.7 Sequencing prepared tagged amplicons on the Illumina HiSeq2000

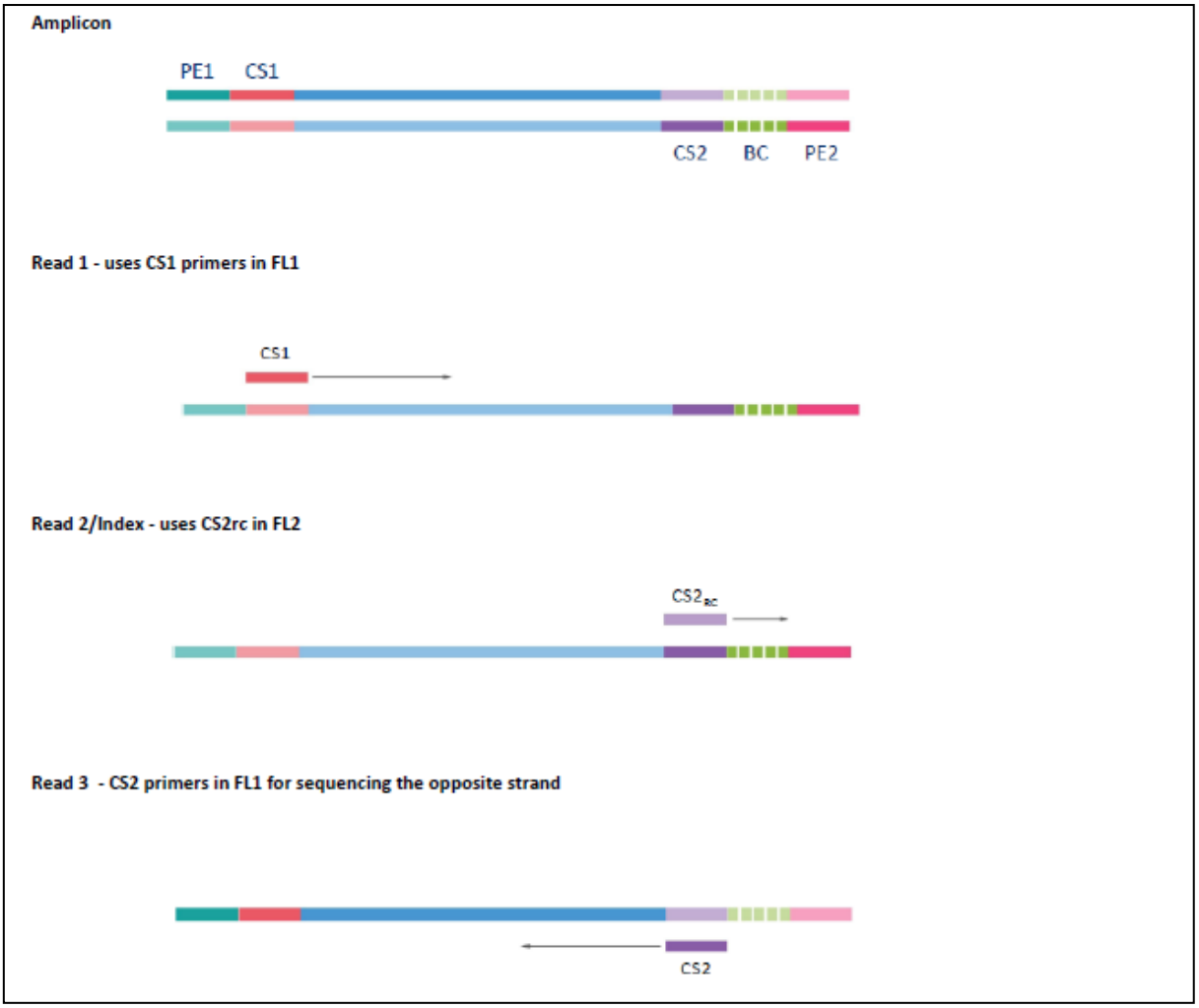


Figure 6.7 Sequencing prepared tagged amplicons on the Illumina HiSeq2000. *PE1=paired end adaptor, PE2=paired end adaptor 2, CS1=custom sequence tag 1, CS2=custom sequence tag 2, CS2rc=custom sequencing primer 2 reverse complement, BC=barcode.*

The diagram (Figure 6.8) overleaf describes the sequencing method further by adding details at the base pair level (bp).

Figure 6.8 The sequencing method at the base pair (bp) level

dsDNA construct (amplicon)

PE1- acactgacgacatggttctaca[insert]-agaccaagtctctgcaccgta-[Index]-PE2

PE1- tgtgactgctgtaccaagatgt[insert]-tacggtagcagagacttggtct-[Index]-PE2

First read (forward)

PE1-acactgacgacatggttctaca[insert]-agaccaagtctctgcaccgta-[Index]-PE2

Second (Index) read

PE1- acactgacgacatggttctaca[insert]-agaccaagtctctgcaccgta-[Index]-PE2

Third read (reverse)

PE1-tgtgactgctgtaccaagatgt[insert]-tacggtagcagagacttggtct-[Index]-PE2

Figure 6.8 The sequencing method at the base pair (bp) level CS1 = custom sequencing primer 1, CS2= custom sequencing primer 2, CS2rc = custom sequencing primer 2 reverse complement. *This diagram shows at the base pair level how sequencing is achieved.*

6.5.2 Preparation of sequencing reagents

Illumina TruSeq Reagents are used with the custom sequencing primers from Fluidigm (FL1 and FL2).

FL1 is the sequencing primer and contains 50μM each of CS1 and CS2 primers (common sequence tags).

FL2 is the barcode primer and contains 50μM each of the CS1rc and CS2rc primers

Table 6.9 The sequences of CS1/CS2 primers.

Primer	Sequence
CS1	5'-ACACTGACGACATGGTTCTACA-3'
CS2	5'-TACGGTAGCAGAGACTTGGTCT-3'
CS1rc	5'-TGTAGAACCATGTCGTCAGTGT-3'
CS2rc	5'-AGACCAAGTCTCTGCTACCGTA-3'

Table 6.9 The sequences of CS1/CS2 primers.

Table 6.10 Preparation of reagents for read one (forward)

Reagent	Volume
TruSeq reagent HP6	990 μl
FL1	10 μl
Total	1000 μl

Table 6.10 Preparation of reagents for read one (forward)

FL1 reagent is diluted according to the Table 6.10 above at a final concentration of 0.50 μM in Illumna TruSeq reagent HP6 in a DNAse, RNAse free 0.5ml microcentrifuge tube. The tube is vortexed for 30 seconds.

Table 6.11 Preparation of reagents for read two (index)

Reagent	Volume
TruSeq reagent HP8	990 μl
FL2	10 μl
Total	1000 μl

Table 6.11 Preparation of reagents for read two (index)

The FL2 reagent is diluted according to the Table 6.11 above at a final concentration of 0.50 µM in Illumina TruSeq reagent HP8 in a DNAse, RNAse free 0.5ml microcentrifuge tube. The tube is vortexed for 30 seconds.

Table 6.12 Preparation of reagents for read three (reverse)

Reagent	Volume
TruSeq reagent HP7	990 µl
FL1	10 µl
Total	1000 µl

Table 6.12 Preparation of reagents for read three (Reverse)

The FL1 reagent is diluted according to the Table 6.12 above at a final concentration of 0.50 µM in Illumina TruSeq reagent HP7 in a DNAse, RNAse free 0.5ml microcentrifuge tube. The tube is then vortexed for 30 seconds.

6.5.3 Cluster generation

Cluster generation is performed on the cBot, which is a separate instrument from the HiSeq. This step is performed following Illumina protocols for HiSeq2000 and is essentially a 'plug and play' process using pre-prepared Illumina reagents in sealed 96 well plate. This reduces any possibility of sample contamination and almost eliminates the hands on time for cluster generation. Clustering takes 4 hours.

First, the 96 well reagent plate is prepared by thawing and vortexing. The foil over each tube in well 10 is pierced. Then libraries are diluted to 20 pM with NaOH to denature; the libraries are then diluted to 10 pM with Illumina HT1 hybridisation buffer and 120 µl is placed into an 1 tube of an 8 tube strip (each tube represents 1 lane on the flow cell). A pre-wash is performed following the instructions on the screen. The complete cluster generation run protocol is then selected (i.e. this performs amplification, linearisation, blocking and primer hybridisation). The 96 well reagent plate is then loaded and following the steps on the screen. The flow cell is then carefully slid into the path of the barcode scanner to be read then washed in deionised water and dried. The flow cell is positioned onto the thermal stage in the cBot and the manifold loaded by positioning it over the flow cell and aligning it by the guide pins located on the thermal stage. The manifold is then locked into position. The outlet end of the manifold is connected to the outlet port and the rear clamp secured. The sipper combs are aligned and secured into place. The libraries in the 8-tube strip are loaded

into the ‘Template’ row on the cBot and finally the primers are loaded into the ‘Primer’ row. A pre-run check is performed following on screen instructions and the run started.

6.5.4 Sequencing on the HiSeq2000

The prepared flow cell is now ready for loading into the HiSeq for sequencing. The pre-prepared sequencing and indexing reagents are loaded into the HiSeq following the instructions on the screen on the instrument. The reagents are loaded in the appropriate rack position inside the instrument according to the run being performed (i.e a dual indexed run on a paired-end flow cell). The caps are removed from the reagent tubes and slid into position and aligned. After closing the reagent door the reagents are set to ‘prime’. The flow cell is then removed from the cBot and positioned in the HiSeq2000 by positioning the flow cell lever to position 1 and allowing the vacuum to engage the flow cell. After 5 seconds the flow cell lever is moved to position 2 and the sequencing run is started.

6.5.5 Bioinformatics and data analysis

Figure 6.9 An overview of the bioinformatics and data analysis

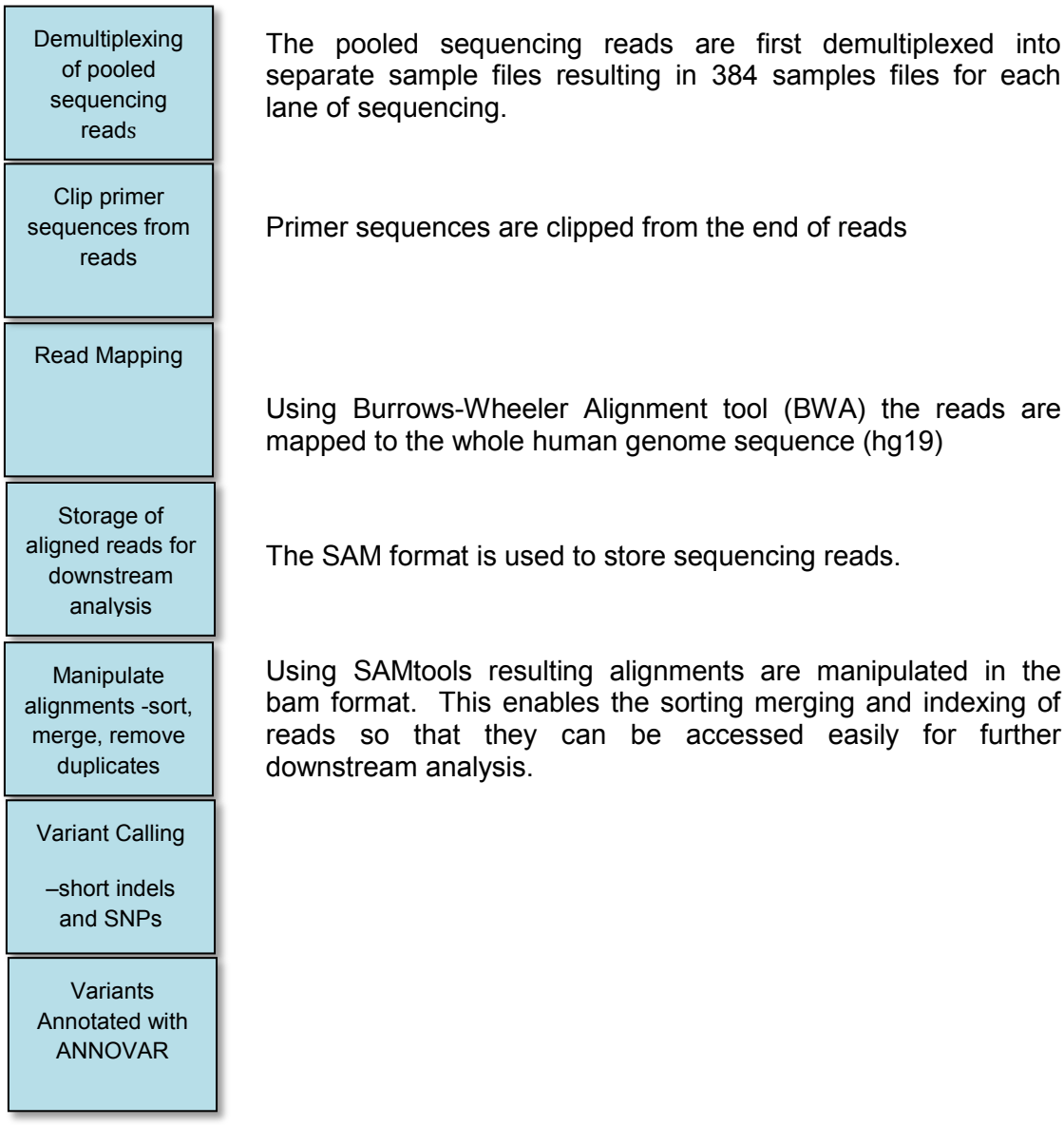


Figure 6.9 Overview of the bioinformatics and data analysis

Figure 6.9 shows an overview of the bioinformatics and data analysis performed by Ed Dicks at Strangeways Research Laboratory, Cambridge UK.

6.5.5.1 Demultiplexing

Demultiplexing of pooled sequencing reads is performed by Bioinformaticians at Source Bioscience Plc., Nottingham, UK. This had to be conducted in four batches of 96 samples for each sequencing lane.

6.5.5.2 Read mapping with BWA.

BWA (Li & Durbin 2009) is the chosen as the read alignment program and is performed by Ed Dicks, Strangeways Research Laboratory, Cambridge, UK. However, I tried this myself using the script presented here in Figure 6.10. This program is fast and scalable for large numbers of samples and has the advantage of a smaller memory requirement compared to other read mapping programs. BWA allows for gapped alignment of reads and this is essential in mapping reads containing indels. The BWA output is in standard SAM (Sequence Alignment/Map) format. This allows for the resulting sequencing files to use the SAMtools package for downstream analysis including variant calling.

BWA is first downloaded from http://sourceforge.net/projects/bio-bwa/files/. Then the files are first unzipped. Using the terminal the program required building using unix commands. Directories are created so that BWA is accessible and placed in the correct PATH.

The latest build of the whole human genome sequence was downloaded hg19 using the bundle provided by the Genome Anaysis Toolkit (GATK)

http://www.broadinstitute.org/gsa/wiki/index.php/GATK_resource_bundle and this is used as the reference sequence.

A bash shell script is written to perform the alignment against the whole human genome sequence in a loop to include all samples in order.

Figure 6.10 Bash shell script written to perform the alignment against the whole human genome.

```
#!/bin/bash

#for loop run in Project_Jane_Hayward directory
#find max depth 1 only directories
for i in $(find /Users/janehayward/Documents/Project_Jane_Hayward -mindepth 1 -maxdepth 1 -type d); do
echo $i
cd "$i"
#strip ./Sample_ from directory name
sample=$(echo $i | awk -F'/' '{print $NF}')
PU=$(head -n 1 L007_R1.fq | awk -F':' '{print $NF}')

echo $sample $PU
echo
```

```
#reference
ref="/Users/janehayward/NGSapps/hg19/ucsc.hg19.fasta"

# t = threads for quad core mac
#bwa aln -t 3 "$ref" L007_R1.fq > L007_R1.sai
#bwa aln -t 3 "$ref" L007_R2.fq > L007_R2.sai
bwa aln "$ref" L007_R1.fq > L007_R1.sai
bwa aln "$ref" L007_R2.fq > L007_R2.sai
bwa sampe "$ref" L007_R1.sai L007_R2.sai L007_R1.fq L007_R2.fq > "$sample".sam

##Convert sam to bam and add read groups
java -Xmx2g -jar ~/NGSapps/bin/AddOrReplaceReadGroups.jar \
I="$sample".sam  \
O="$sample".sorted.RG.bam \
LB=Batch1 \
PL=illumina \
PU="$PU" \
SM="$sample" \
SORT_ORDER=coordinate \
CREATE_INDEX=true

cd ../

done
```

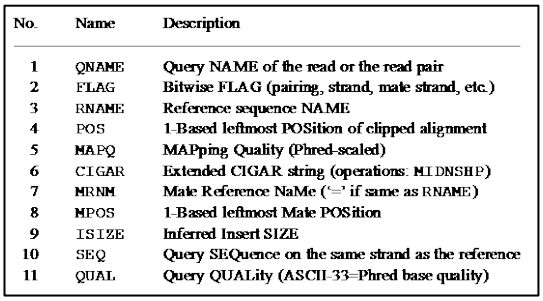
Figure 6.10 Bash shell script written to perform the alignment against the whole human genome. *This is the script I used to learn how to perform read alignment for the first lane sequenced.* ***Daniel Leongamornlert at Institute of Cancer Research assisted with writing this script***

This script gave sorted BAM files as output. These files sorted against the indexed hg19.fasta could then be viewed using Integrative Genome Viewer (IGV).

6.5.5.3 Manipulate alignments using SAMtools

SAM tools is downloaded from http://sourceforge.net/projects/samtools/files/ and the files first unzipped and using the terminal the package SAM tools is built using unix commands. Again directories are created so that SAMtools is accessible and placed in the correct PATH. SAMtools is chosen because the standard format is the SAM format; it is scalable and suitable for large sets of alignment files. The SAM format is a tab delimited format with a header section and an alignment section. The alignment section includes lines with 11 mandatory fields and variable optional fields. The image (Figure 6.11) below shows those mandatory fields.

Figure 6.11 SAM tools mandatory fields



No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Figure 6.11 SAM tools mandatory fields. *There are 11 mandatory fields in the SAM tools package*

6.5.5.4 Storage of reads for downstream analysis

Raw sequencing files are converted to SAM format for storage of reads and accessibility of reads for downstream analysis.

6.5.5.5 Manipulate alignments

Manipulating alignments involves sorting, merging and indexing. These manipulations are performed with SAM tools.

6.5.5.6 Variant calling

Variant calling (SNP and Indel discovery) is achieved with Genome Analysis Tool Kit, GATK (McKenna et al 2010). These are used for base quality score recalibration, local indel realignment, and SNP and INDEL discovery and genotyping across all 2,636 samples simultaneously using standard hard filtering parameters or variant quality score recalibration

GATK software and recommended practice for bam recalibration and read/base filtering are performed. Quality score recalibration is conducted, however, local realignment and duplicate removal was not required as sequencing Fluidigm Access Array amplicons means that sequencing always starts from the same genomic position and at high depth coverage per exon; this compares to sequencing of amplicons generated via alternative enrichment technology (such as long range PCR and genomic capture) in which sequencing would begin from random points within amplicons so that sequences containing indels may require local realignment as in the pilot study. All variants are checked in NCBI dbSNP database and where known the rs numbers are assigned to each.

6.5.5.7 Variant annotation

The final variants are annotated using ANNOVAR (Wang et al 2010). This software

tool was designed by Wang et al (2010) to identify biologically important genetic variants. It annotates single nucleotide variants as well as indels by assessing the significance of these on gene function. This is done by calculating functional importance scores, distinguishing those variants previously found in 1000 Genomes Project data and on dbSNP

6.5.5.8 Final filtering of variants

This is performed to eliminate possible sequencing artefacts from real variants. The following parameters are used to filter out spurious variant calls:

- Variants detected are accepted if read depth is 15 X or more and the alternate allele frequency is 40% or more; or if the read depth is 30 X or more and the alternate allele frequency is 30% or more.

If variants are detected in regions that did not meet these criteria then variants are rejected.

6.5.5.9 Predicting which missense changes are deleterious

Two programs PROVEAN and PolyPhen-2 are used to predict the functional effect of novel missense changes. PROVEAN is chosen as it uses a substitution matrix, including 20 amino acid residues, which assesses the similarity between the subject sequence and the reference sequence. PolyPhen-2 is chosen as an alternative comparison tool to evaluate concordance between the two programs and to ensure that those missense variants that are likely to be pathogenic are not missed.

6.6 Methodology for Chapter Four

Laboratory work was conducted myself in UCL and in the Great Ormond Street Molecular Genetics Laboratory. Maria Intermaggio and Andre Kim also conducted laboratory work at University of Southern California (USC) Bioinformatics was conducted by Christopher K Edlund at USC.

The same NGS analysis is conducted as described in the previous section detailing the methodology for chapter 3. This study started by including 10 candidate genes, however, 1 gene (TiPARP) is removed from the analysis, as this does not elicit any data. These remaining 9 genes are both known and unknown ovarian cancer susceptibility genes and are sequenced in 2,300 women that are part of a clinical screening trial for ovarian cancer, the UK Familial Ovarian Cancer Screening Study (UKFOCSS).

6.6.1 DNA Samples

6.6.1.1 UK Familial Ovarian Cancer Screening Study (UK FOCSS)

This study recruits women with an increased risk of developing ovarian cancer due to family history or an inherited predisposition to the disease. Screening takes the form of an annual CA125 blood test (a tumour marker) alongside annual transvaginal ultrasound scanning of the ovaries (Figure 6.12). Later this would be increased to 3 screening visits per year for blood and ultrasound tests.

The primary aim of UK FOCSS is to establish an approach for screening for ovarian cancer. This approach requires examining parameters such as screening intervals, types of screening test and also includes assessments of cost and morbidity. In addition the study creates a serum bank, which will be invaluable in researching novel tumour markers.

All of the women sequenced in this 9 gene candidate study are in the UKFOCSS study and therefore, were at a significantly increased risk of developing ovarian cancer.

Figure 6.12 UK FOCSS study design

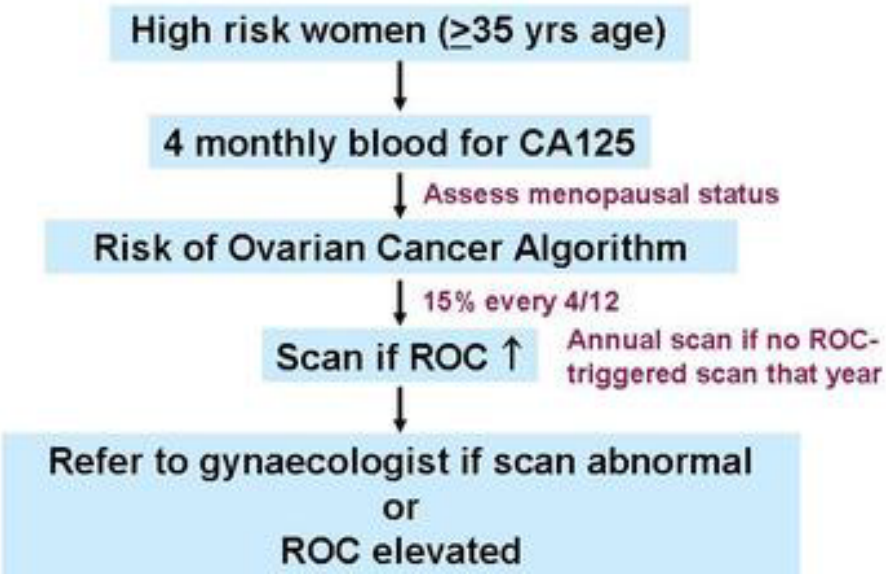


Figure 6.12 UK FOCSS study design. *This flow diagram describes the UK FOCSS study design investigating screening strategy for high-risk women*

6.6.2 Target enrichment

The study sequences the coding regions only of known ovarian cancer susceptibility genes (*BRCA1*, *BRCA2*, *RAD51C* and *RAD51D*) and novel candidate genes *BRIP1*, *PALB2*, *NBN*, *BARD1*, and *RAD51B*.

6.6.2.1 Primer design

RAD51B, *RAD51C*, *RAD51D* primers are used from the previous study (Chapter 3 Methodology) and designed by the Fluidigm Primer Design service. *BRCA1*, *BRCA2* primers are designed by Honglin Song from Strangeways Research Laboratory in Cambridge University. Primers for candidate genes *BRIP1*, *PALB2*, *BARD1* and *NBN* are designed by myself using the NCBI Primer BLAST tool (http://www.ncbi.nlm.nih.gov/tools/primer-blast/).

6.6.2.2 Primer multiplexing and pooling

Primers are multiplexed to a maximum of 10 amplicons per well on the Fluidigm Access Array platform meaning that PCR reactions amplifying all coding regions of all 9 genes are conducted on one Access Array chip.

The forward and reverse primer pairs are pooled following the recommended criteria suggested in the Fluidigm Access Array protocols as follows:

1. Primer pairs that produce overlapping PCR products are not mixed in the same well. This is especially relevant where more than one primer set is required to cover an exon in which case these primer sets are positioned in separate wells.
2. Primer sets are amplifying regions that are at least 5Kb apart when mixed in the same well.
3. Primer sets are mixed that are within an amplicon size range of 20% of the average size of the pool. Generally, amplicons between 140bp and 210bp were aimed for.
4. Primer pairs with a similar GC % content were mixed in the same well.
5. Primers are designed to be specific to the desired regions, i.e. they do not have multiple annealing sites within the sample DNA template; this is verified by using NCBI BLAST on line tool to ensure that targets are specific.

6.6.2.3 Pooling of primer pairs on the Fluidigm Access Array

48 pools were created from 406 amplicons the tables in Appendix X describe the location of each primer pair.

6.6.3 Library preparation

6.6.3.1 Multiplex amplicon tagging for Illumina on the 48.48 Access Array IFC

Sequencer specific tagged amplicons are generated in two PCR reactions as in the previous study. Each reaction is multiplexed at a maximum of 10 pairs of primers in each well. 406 amplicons are created on one Access Array Integrated Fluidic Circuit (IFC). Following the first PCR reaction, the products are harvested and then used as a template in the second PCR reaction, during which the 384-barcode sequences are added along with the Illumina sequencing adaptors. Each IFC produces 406 amplicons in each of 48 samples, each one uniquely indexed to allow for the pooling of all 19,488 PCR products.

Table 6.13 Preparation of 20X primer solutions

Component	Volume (µl)	Final Concentration
CS1-TS-F (50uM)	6.0 per primer	1µM
CS2-TS-R (50uM)	6.0 per primer	1µM
20X Access Array loading reagent	15.0	1X
DNA Suspension Buffer (TE)	165.0	
Total	300.0	

Table 6.13 Preparation of 20X primer solutions

Where there are less than 10 primer pairs in the well additional DNA suspension buffer is added to maintain correct primer concentration.

6.6.3.2 Preparation of 20X primer solutions

20X primer solutions for 406 primer pairs are prepared. The Table 6.13 above shows the primer dilution for one well (of 48 wells) with 10 primer pairs.

In a 96 well plate 12µl of each pooled forward and reverse primer pair are added at a mixture of up to 10 primer pairs in each well. This produces enough for 60 Access Arrays. The 20X primer solutions are divided into 5µl aliquots. This means that each 5µl aliquot is ready prepared for chip set up and to eliminate continued freeze/thaw cycles. Only 4µl is loaded onto the access array, but 5µl prepared to allow for dead volume.

The set-up of the Fluidigm Access Array system is described previously. Access array chips are prepared in an identical manner as for Chapter 3 Methodology for the case-control study.

6.6.3.3 Priming, set up and running the 48.48 access array

Priming the chips, setting up and running of each Access Array is performed in an identical manner to that in the former study and is previously described.

6.6.3.4 Checking the barcoded PCR Products

PCR products are checked using an Agilent Bioanalyzer 2100 as described in the previous study.

6.6.3.5 Pooling and purification of PCR products for each Access Array

Pooling and purification of PCR products for each Access Array is conducted in an identical manner to that described for the previous study.

6.6.3.6 Quantitation and normalisation of pools

Quantitation and normalisation of pooled PCR products is conducted in an identical manner to the previous study and is already described.

6.6.3.7 Dilution of pools to 10nM

Dilution of pools is achieved in an identical manner to that previously described for the case-control study.

6.6.3.8 Final quality control check

A final QC check is performed. Each diluted pool is quantified in triplicate using the Agilent Bioanalyzer 2100. The actual final concentration is given as the mean across the triplicate repeats.

6.6.4 Sequencing

6.6.4.1 Sequencing prepared tagged amplicons on the Illumina HiSeq2000.

Preparation of sequencing reagents (Illumina TruSeq) and sequencing on the Illumina HiSeq2000 are performed in an identical manner to that in the previous study.

6.6.4.2 Cluster generation

Clustering is performed on the CBot following Illumina protocols for HiSeq2000 described in detail for the case-control study.

6.6.5 Bioinformatics and data Analysis

The Bioinformatics and data analysis are performed by Scientists at University of Southern California (USC) as follows, information provided by Christopher K Edlund (5-9-2013)

Sequencing reads are generated by the USC Epigenome Center using the Illumina HiSeq2000 instrument. A total of six lanes are run using 384-barcoded samples pooled together in each lane.

Sequencing reads are de-multiplexed and aligned to the hg19 genome (updated chrM, randoms included, haps removed) using BWA version 0.6.1-r104, resulting in one BAM file per sample.

Reads mapping outside the targeted regions are removed using the intersectBed tool in the bedTools software.

Reads are aligned around indels using the GATK IndelRealigner tool and known indels used for realignment are obtained from the following VCF files: Mills_and 1000G_gold_standard.indels.hg19.vcf and 1000G_phase1.indels.hg19.vcf.

For each read starting inside a primer, all bases overlapping and on the same strand as the primer are soft-clipped.

Samples are merged together by lane for running the GATK BaseRecalibrator too available:http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_bqsr_BaseRecalibrator.html. A recalibration table is generated for each lane using the default covariates in GATK. Known variants sites used in this process are obtained from the following VCF files: Mills_and_1000G_gold_standard.indels.hg19.vcf, 1000G_phase1.indels.hg19.vcf, and dbsnp_137.hg19.chrMT.vcf. Base quality scores are recalibrated for each sample using the recalibration table. Samples are then unmerged.

The GATK ReduceReads tool is used to compress each BAM file. Default settings are used.

The GATK UnifiedGenotyper tool is used to call variants with all reduced BAM files as input. The following settings are used:

--dbsnp $GATK/resources/dbsnp_137.hg19.chrMT.vcf

-stand_call_conf 30.0

-stand_emit_conf 30.0

-dcov 1000

--min_base_quality_score 20

--output_mode EMIT_VARIANTS_ONLY

--genotype_likelihoods_model BOTH

References

Aarnio M, Sankila R, Pukkala E, Salovaara R, Aaltonen LA, de la Chapelle A, Peltomäki P, Mecklin JP, Järvinen HJ. (1999) Cancer risks in mutation carriers of DNA mismatch-repair genes. Int. J. Cancer. 12;81(2):214-8

Alsop K, Fereday S, Meldrum C, deFazio A, Emmanuel C, George J, Dobrovic A, Birrer MJ, Webb PM, Stewart C, Friedlander M, Fox S, Bowtell D, and Mitchell G. (2012) BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. J Clin Oncol. Vol. 30(21):2654-63.

D'Andrea, A.D and Grompe, M. (2003) The Fanconi anaemia/BRCA pathway. Nat Rev Cancer. Vol.3 (1):**23-34**

Antoniou, A.C. and Easton, D.F (2003) Polygenic Inheritance of Breast Cancer: Implications for Design of Association Studies. Genetic Epidemiology. Vol 25:**190–202**

Antoniou, A.C., Gayther, S.A., Stratton, J.F., Ponder, B.A., Easton, D.F. (2000). Risk models for familial ovarian and breast cancer. Genet Epidemiol 18: 173-190

Antoniou, A., Pharoah, P.D.P., Narod S., Risch, H.A, Eyfjord J.E, Hopper, J.L, Loman, N., Olsson, H., Johannsson, O., Borg, Å., Pasini, B., Radice, P., Manoukian, S. Eccles, D.M., Tang, N., Olah, E., Anton-Culver, H. Warner, E., Lubinski, J. Gronwald15., J. Gorski, B,. Tulinius, H, Thorlacius, S., Eerola, H, Nevanlinna, H., Syrjäkoski, K., Kallioniemi, O.-P, Thompson, D., Evans, C, Peto, J., Lalloo, F, Evans, D.G. and. Easton D.F (2003) Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. AJHG. Vol 72 (5): 1117-1130

Antoniou, A.C., PPD Pharoah, P.D.P., Smith, P. and Easton, D.F. (2004) The BOADICEA model of genetic susceptibility to breast and ovarian cancer British Journal of Cancer (2004) 91, **1580 – 1590**

Antoniou, A.C., Cunningham, A.P., Peto, J., Evans, D.G., Lalloo, F., Narod, S.A. Risch, H.A., Eyfjord, J.E., Hopper, J.L., Southey, M.C., Olsson, H., Johannsson, O., Borg, A., Pasini, B., Radice, P., Manoukian, S., Eccles, D.M., Tang, N., Olah, E., Anton-Culver, H., Warner, E., Lubinski, J., Gronwald, J., Gorski, B., Tryggvadottir, L., Syrjakoski, K., Kallioniemi, O.P., Eerola, H., Nevanlinna, H., Pharoah, P.D., Easton, D.F. (2008) The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br J Cancer. 98(8): **1457-66**

Ashley, D.J.B (1969) The two “hit” and multiple “hit” theories of carcinogenesis. Br J Cancer. Vol 23(2): **313-328**

Ashworth, A. (2008) A synthetic lethal therapeutic approach: poly(ADP)ribose polymerase inhibitors for the treatment of cancers deficient in DNA double-strand beak repair. J Clin Oncol 26: **3785-3790**

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, and Shendure J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. Vol. 12(11):745-55

Barnes, D.R and Antoniou, A.C. (2012) Unravelling modifiers of breast and ovarian cancer risk for BRCA1 and BRCA2 mutation carriers: update on genetic modifiers. J Intern Med. 271(4):**331-43**

Barrow E, Robinson L, Alduaij W, Shenton A, Clancy T, Lalloo F, Hill J, Evans DG. (2009) Cumulative lifetime incidence of extracolonic cancers in Lynch syndrome: a report of 121 families with proven mutations. Clin Genet. 75(2):**141-9**

Bast, R.C. Jr, Feeney, M., Lazarus, H., Nadler, L.M., Colvin, R.B., and Knapp RC.(1981) Reactivity of a monoclonal antibody with human ovarian carcinoma. J Clin Invest. Vol. 68(5):**1331-7**

Bell, D.A. (2005) Origins and molecular pathology of ovarian cancer
Modern Pathology 18, **S19–S32**

Bernstein, C., Bernstein, H., Payne, C.M. and Garewal, H. (2002) DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis. Mutat Res. Vol. 511(2):**145-78**

Bolton KL, Chenevix- Trench G, Goh C, et al. Association Between BRCA1 and BRCA2 Mutations and Survival in Women With Invasive Epithelial Ovarian Cancer. *JAMA.* 2012;307(4):382-389

Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013) Rare disease genetics in the era of next generation sequencing: discovery to translation. Nature Rev. Genet. 2013/09/03/online 10.1038/nrg3555 pp1-11

Braybrooke, J. P., Spink, K. G., Thacker, J., Hickson, I. D. (2000) The RAD51 family member, RAD51L3, is a DNA-stimulated ATPase that forms a complex with XRCC2. J. Biol. Chem. 275: **29100-29106**

Brown, M. A., Nicolai, H., Xu, C.-F., Griffiths, B. L., Jones, K. A., Solomon, E., Hosking, L., Trowsdale, J., Black, D. M., and McFarlane, R. (1996) Regulation of *BRCA1*. (Letter) Nature. Vol 372: **733**

Bunyan, D.J., Eccles, D.M., Sillibourne, J., Wilkins, E., Thomas, N.S., Shea-Simonds, J., Duncan, P.J., Curtis, C.E., Robinson, D.O., Harvey, J.F., and Cross, N.C.P (2004) Dosage analysis of cancer predisposition genes by multiplex ligation-dependent probe amplification. British Journal of Cancer. Vol 91, 1155 – 1159

Campbell, S., Goessens, L., Goswamy, R. and Whitehead, M. (1982) Real-time ultrasonography for determination of ovarian morphology and volume. A possible early screening test for ovarian cancer? Lancet. Vol 1. (8269):**425-6**

Cantor, S. B., Bell, D. W., Ganesan, S., Kass, E. M., Drapkin, R., Grossman, S., Wahrer, D. C. R., Sgroi, D. C., Lane, W. S., Haber, D. A., and Livingston, D. M. (2001) BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. Cell 105: 149-160

Cantor SB, and Guillemette S. (2011) Hereditary breast cancer and the BRCA1-associated FANCJ/BACH1/BRIP1. Future Oncol. Vol ;7(2):253-61

Carr IM, Camm N, Taylor GR, Charlton R, Ellard S, Sheridan EG, Markham AF, and Bonthron DT. (2011) GeneScreen: a program for high-throughput mutation detection in DNA sequence electropherograms. J Med Genet. Vol. 48(2):123-30.

Casadei, S., Norquist, B.M., Walsh, T., Stray, S., Mandell, J.B., Lee, M.K., Stamatoyannopoulos, J.A., King, M.C. (2011) Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. Cancer Res. Vol. 15;71(6):**2222-2229**

Cederquist K, Emanuelsson M, Wiklund F, Golovleva I, Palmqvist R, Grönberg H (2005) Two Swedish founder MSH6 mutations, one nonsense and one missense, conferring high cumulative risk of Lynch syndrome. Clin Genet (6):**533-41**.

Chenevix-Trench, G., Milne, R.L., Antoniou, A.C., Couch, F.J., Easton, D.F., and Goldgar, D.E., on behalf of CIMBA An international initiative to identify genetic modifiers of cancer risk in *BRCA1* and *BRCA2* mutation carriers: the Consortium of Investigators of Modifiers of *BRCA1* and *BRCA2* (CIMBA) (2007) Breast Cancer Research. 9:104

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7(10)

Cirulli, E.T. & Goldstein, D.B. (2010) Uncovering the roles of rare variants in common disease through whole genome sequencing. Nature Rev Gen. vol. 11 **415-425**

Coulet F, Fajac A, Colas C, Eyries M, Dion-Minière A, Rouzier R, Uzan S, Lefranc JP, Carbonnel M, Cornelis F, Cortez A, and Soubrier F. (2012) Germline RAD51C mutations in ovarian cancer susceptibility. Clin Genet. Vol. 83(4):**332-6**

Daly, A.K. & Day, C.P. (2001) Candidate gene case-control association studies: advantages and potential pitfalls. Br J Clin Pharmacol, 52, **489-499**

Danoy, P., Sonoda, E., Lathrop, M., Takeda, S., Matsuda, F. (2007). A naturally occurring genetic variant of human XRCC2 (R188H) confers increased resistance to cisplatin-induced DNA damage. Biochem. Biophys. Res. Commun. 19; 352 (3):763-8

Davidson CJ, Zeringer E, Champion KJ, Gauthier MP, Wang F, Boonyaratanakornkit J, Jones JR, Schreiber E. (2012) Improving the limit of detection for Sanger sequencing: A comparison of methodologies for KRAS variant detection. BioTechniques. Vol. 53, No. 3, **182-188**

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl., C., Philippakis, A., del Angel, G., Rivas, M.A, Hanna, M., McKenna, A., Fennell, T. Kernytsky, A., Sivachenko, A, Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 43(5):**491-498**.

Fackenthal, J.D. and Olopade, O.I. (2007) Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. Nature Reviews Cancer 7, 937-948

Domchek SM, Friebel TM, Singer CF, Evans DG, Lynch HT, Isaacs C, Garber JE, Neuhausen SL, Matloff E, Eeles R, Pichert G, Van t'veer L, Tung N, Weitzel JN, Couch FJ, Rubinstein WS, Ganz PA, Daly MB, Olopade OI, Tomlinson G, Schildkraut J, Blum JL, Rebbeck TR. (2010) Association of risk-reducing surgery in *BRCA1* or *BRCA2* mutation carriers with cancer risk and mortality. JAMA vol 304(9): **967-975**

Dubeau, L (2008) The cell of origin of ovarian epithelial tumours. The Lancet Oncology. Vol 9 (12): **1191-1197**

Douma KF, Aaronson NK, Vasen HF, Verhoef S, Gundy CM, and Bleiker EM. (2010) Attitudes toward genetic testing in childhood and reproductive decision-making for familial adenomatous polyposis. Eur J Hum Genet. Vol. 18(2):186-93.

Easton, D.F., Bishop, D.T., Ford, D., Crockford, G.P., the Breast Cancer Linkage Consortium (1993). Genetic linkage analysis in familial breast and ovarian cancer. Am J Human Genet 52: **678-701**

Engel, C., Versmold, B., Wappenschmidt, B., Simard, J., Easton, D.F., Peock, S., Cook, M., Oliver, C., Frost, D., Mayes, R., Evans, D.G., Eeles, R., Paterson, J., Brewer, C.; for the Epidemiological Study of Familial Breast Cancer (EMBRACE), McGuffog, L., Antoniou, A.C., Stoppa-Lyonnet, D., Sinilnikova, O.M., Barjhoux, L., Frenay, M., Michel, C., Leroux, D., Dreyfus, H., Toulas, C., Gladieff, L., Uhrhammer, N., Bignon, Y-J., Meindl, A., Arnold, N., Varon-Mateeva, R., Niederacher, D., Preisler-Adams, S., Kast, K., Deissler, H., Sutter, C., Gadzicki, D., Chevenix-Trench, G., Spurdle, A.B., Chen, X., Beesley, J.; for the Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer (kConFab), Olsson, H., Kristofferson, U., Ehrencrona, H., Liljegren, A.; for the Swedish Breast Cancer Study, Sweden (SWE-BRCA), van der Luijt, R.B., van Os, T.A., van Leeuwen, F.E.; for the Hereditary Breast *and* Ovarian cancer group Netherlands (HBON), Domcheck, S.M., Rebbeck, T.R., Nathanson, K.L., Osorio, A., y Cajal, T.R., Konstantopolou, I., Benifez, J., Friedman, E., Kaufman, B., Laitman, Y., Mai, P.L., Greene, M.H., Nevanlinaa, H., Aittomaki, K., Szabo, C.I., Caldes, T., Couch, F.J., Andrulis, I.L., Godwin, A.K., Hamann, U. and Schmutzler, R.K.; on behalf of the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) (2010). Association of the Variants CASP8 D302H and CASP10 V410I with Breast and Ovarian Cancer Risk in *BRCA1* and *BRCA2* Mutation Carriers. Cancer Epidemiol Biomarkers Prev 19; **2859**

Erkko, H., Xia, B., Nikkila, J., Schleutker, J., Syrjakoski, K., Mannermaa, A., Kallioniemi, A., Pylkas, K., Karppinen, S.-M., Rapakko, K., Miron, A., Sheng, Q., and 15 others. (2007) A recurrent mutation in PALB2 in Finnish cancer families. Nature 446: **316-319**

Evans, D.G., Young, K., Bulman, M., Shenton, A., Wallace, A. and Lalloo, F. (2008) Probability of BRCA1/2 mutation varies with ovarian histology: results from screening 442 ovarian cancer families. Clin Genet. Vol. 73(4):**338-45**.

Fackenthal, J, D. and Olopade, O.I. (2007) Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. Nature Reviews Cancer 7, **937-948**

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M. Dillon, K.J., Hickson, I., Knights, C., Martin, N.M.B., Jackson, S.P., Smith, G.C.M., and Ashworth, A. (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature. Vol: 434: 917-921

Farrell, P.M., Rosenstein, B.J., White, T.B., Accurso, F.J., Castellani, C., Cutting, G.R., Durie, P.R., Legrys, V.A., Massie, J., Parad, R.B., Rock, M.J., Campbell, P.W. 3rd; Cystic Fibrosis Foundation (2008). Guidelines for diagnosis of cystic fibrosis in newborns through older adults: Cystic Fibrosis Foundation consensus report. J Paediatric Medicine. Vol.153(2):S4-S14.

Ferla, R., Calo, V., Cascio, S., Ranaldi, G., Badalamenti, G., Carreca, I., Surmacz, E., Goloucci, G. Bazan, V. and Russo, A (2007) Founder mutations in BRCA1 and BRCA2 genes. Annals of Oncology 18 (Supplement 6); **vi93-vi98**.

FIGO Committee on Gynecologic Oncology (2009) Current FIGO staging for cancer of the vagina, fallopian tube, ovary, and gestational trophoblastic neoplasia. Int J Gynaecol Obstet 105 (1): 3-4

Fletcher, O and Houlston, R.S. (2010) Architecture of inherited susceptibility to common cancer. Nat Rev Cancer Vol 10:**353-361**

Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude J, Sobol H, Teare MD, Struewing J, Arason A, Scherneck S, Peto J, Rebbeck TR, Tonin P, Neuhausen S, Barkardottir R, Eyfjord J, Lynch H, Ponder BA, Gayther SA, Zelada-Hedman M, and the Breast Cancer Linkage Consortium. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet 62:676–689.

Foulkes, A.S., (2009) Applied statistical genetics with R. For population-based association studies. Springer.
Gage, M., Wattendorf, D. and Henry, LR. (2012) Translational advances regarding hereditary breast cancer syndromes. J Surg Oncol. Vol. 105(5):**444-51**

de Garibay GR, Díaz A, Gaviña B, Romero A, Garre P, Vega A, Blanco A, Tosar A, Díez O, Pérez-Segura P, Díaz-Rubio E, Caldés T, de la Hoya M. (2012) Low prevalence of SLX4 loss-of-function mutations in non-BRCA1/2 breast and/or ovarian cancer families. European Journal of Human Genetics 1- 4

Gayther, S.A. and Pharoah, P.D.P (2010) The inherited genetics of ovarian and endometrial cancer. Curr Opin Genet Dev. Vol. 20(3): 231–238

Gayther, S.A (2012) Inherited risk of ovarian cancer and the implications for screening. Intl J Gynaecol Cancer. Vol 22 S12-S15

Gayther, S.A., Mangion, J., Russell, P., Seal, S., Barfoot, R., Ponder, B.A.J., Stratton, M.R., & Easton, D (1997) Variation of risks of breast and ovarian cancer associated with different germline mutations of the BRCA2 gene. Nature Genetics 15, 103 – 105

Gayther SA, Russell P, Harrington P, Antoniou AC, Easton DF, Ponder BA. (1999). The contribution of germline BRCA1 and BRCA2 mutations to familial ovarian cancer: no evidence for other ovarian cancer-susceptibility genes. Am J Hum Genet 65:1021–1029

Gilks, C.B, and Prat, J., (2009) Ovarian carcinoma pathology and genetics: recent advances. Human Pathology Vol. 40, 1213–1223

Goode, E.L, Chenevix-Trench, G., Song, H., Ramus, S.J., Notaridou, M., Lawrenson, K., Widschwendter, M., Vierkant, R.A., Larson, M.C., Kjaer, S.K., Birrer MJ, Berchuck A, Schildkraut J, Tomlinson I, Kiemeney LA, Cook LS, Gronwald J, Garcia-Closas M, Gore ME, Campbell I, Whittemore AS, Sutphen R, Phelan C, Anton-Culver H, Pearce CL, Lambrechts D, Rossing MA, Chang-Claude J, Moysich KB, Goodman MT, Dörk T, Nevanlinna H, Ness RB, Rafnar T, Hogdall C, Hogdall E, Fridley BL, Cunningham JM, Sieh W, McGuire V, Godwin AK, Cramer DW, Hernandez D, Levine D, Lu K, Iversen ES, Palmieri RT, Houlston R, van Altena AM, Aben KK, Massuger LF, Brooks-Wilson A, Kelemen LE, Le ND, Jakubowska A, Lubinski J, Medrek K, Stafford A, Easton DF, Tyrer J, Bolton KL, Harrington P, Eccles D, Chen A, Molina AN, Davila BN, Arango H, Tsai YY, Chen Z, Risch HA, McLaughlin J, Narod SA, Ziogas A, Brewster W, Gentry-

Maharaj A, Menon U, Wu AH, Stram DO, Pike MC; Wellcome Trust Case-Control Consortium, Beesley J, Webb PM; Australian Cancer Study (Ovarian Cancer); Australian Ovarian Cancer Study Group; Ovarian Cancer Association Consortium (OCAC), Chen X, Ekici AB, Thiel FC, Beckmann MW, Yang H, Wentzensen N, Lissowska J, Fasching PA, Despierre E, Amant F, Vergote I, Doherty J, Hein R, Wang-Gohrke S, Lurie G, Carney ME, Thompson PJ, Runnebaum I, Hillemanns P, Dürst M, Antonenkova N, Bogdanova N, Leminen A, Butzow R, Heikkinen T, Stefansson K, Sulem P, Besenbacher S, Sellers TA, Gayther SA, Pharoah PD; Ovarian Cancer Association Consortium (OCAC) (2010) A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nat Genet 42(10): **874-9.**

Górski B, Jakubowska A, Huzarski T, Byrski T, Gronwald J, Grzybowska E, Mackiewicz A, Stawicka M, Bebenek M, Sorokin D, Fiszer-Maliszewska Ł, Haus O, Janiszewska H, Niepsuj S, Góźdź S, Zaremba L, Posmyk M, Płuzańska M, Kilar E, Czudowska D, Waśko B, Miturski R, Kowalczyk JR, Urbański K, Szwiec M, Koc J, Debniak B, Rozmiarek A, Debniak T, Cybulski C, Kowalska E, Tołoczko-Grabarek A, Zajaczek S, Menkiszak J, Medrek K, Masojć B, Mierzejewski M, Narod SA, Lubiński J. (2004) A high proportion of founder BRCA1 mutations in Polish breast cancer families. Int J Cancer.110(5):**683-6**

Greggi, S., Ponder, B.A., Mancuso, S. (1991) Establishment of a European registry for familial ovarian cancer. Eur J Cancer. 27(2):**113–115**

Hanahan, D., & Weinberg, R.A., (2000) The Hallmarks of Cancer. Cell, Vol. 100, 57–70

Harismendy, O., & Frazer, K.A. (2009) Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing by-synthesis technology. BioTechniques Vol. 46:**229-231**

Harismendy, O., Ng, P.C., Strausberg, R. L., Wang, X.W., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S. S., Topol, E.J., Levy, S., and Frazer, K.A. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology. Vol. 10: R32

Harris, T.J and McCormick, F (2010) The molecular pathology of cancer. Nat Rev Clin Oncol. Vol. 7(5):**251-65**.

Harper, J.W., & Elledge, S.J. (2007) The DNA damage response: ten years after. Molecular Cell 28: 739-745

Hawkins, A.K. and Ho, A. (2012) Genetic counseling and the ethical issues around direct to consumer genetic testing. J Genet Counsel. Vol 21: **367-373**

Hellebrand, H., Sutter, C., Honisch, E., Gross, E., Wappenschmidt, B., Schem, C., Deissler, H., Ditsch, N., Gress, V., Kiechle, M., Bartram, C.R., Schmutzler RK, Niederacher D, Arnold N, Meindl A. (2011) Germline mutations in the PALB2 gene are population specific and occur with low frequencies in familial breast cancer. Hum Mutat. 32(6):E2176-88.

Høgdall EV, Christensen L, Kjaer SK, Blaakaer J, Bock JE, Glud E, Nørgaard-Pedersen B, and Høgdall CK.(2003) Distribution of HER-2 overexpression in ovarian carcinoma tissue and its prognostic value in patients with ovarian carcinoma: from the Danish MALOVA Ovarian Cancer Study. Cancer. Vol. 98(1):66-73.

Hollestelle, A., Wasielewski, M., Martens, J.W.M. and Schutte, M. (2010) Discovering moderate-risk breast cancer susceptibility genes. Current Opinion in Genetics & Development. 20:**268–276**

Holschneider, C.H. and Berek, J.S. (2000) Ovarian cancer; epidemiology, biology and prognostic factors. Seminars in Surgical Oncology. Vol. 19 (1): **3–10**

Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2009) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Briefings in Bioinformatics. Vol.11. No 2. 181-197

Huen, M.S.Y., M.H. Sy, S.M.H., and Chen, J. (2010) BRCA1 and its toolbox for the maintenance of genome integrity. Nat. Rev. Molecular Cell Biology. Vol 11:138-148
Jackson, S.P. & Bartek, J. (2009) The DNA-damage response in human biology and disease. Nat. Rev. Vol 461: 1071-1078

Johnson, R. D., Liu, N., Jasin, M. (1999) Mammalian XRCC2 promotes the repair of DNA double-strand breaks by homologous recombination. Nature 401: **397-399**

Jones, S., Hruban, R.H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D.W., Cheng-Ho-Lin, J., Pamisano, E., Brune, K., Jaffee, E.M., Iacobuzio-Donahue, C.A., Maitra, A., Parmigiani, G., Kern, S.E., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Eshleman, J.R., Goggins, M., and Klien, A.P. (2009) Exome sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. Science Vol 324 **pp217**

King, M.C. Joan H. Marks, J.H., and Jessica B. Mandell, J.B. (2003) Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. Science, Vol 302, **643-646**

Kottemann, M.C., Smogorzewska, A. (2013) Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature. Vol. 493 (7432):356-63

Kurman, R.J. and Shih, Ie-M. (2010) The Origin and Pathogenesis of Epithelial Ovarian Cancer: A Proposed Unifying Theory Am J Surg Pathol. Vol 34 (3)

Lander, E.S and Waterman, M.S.(1988) Genomic mapping by fingerprinting random clones: a mathematical analysis, Genomics 2(3): **231-239.**

Lakhani, S.R., Manek, S., Penault-Llorca, F., Flanagan, A., Arnout, L., Merrett, S., McGuffog, L., Steele,D., Devilee,P., Klijn,J.G.M., Meijers-Heijboer,H., Radice,P., Pilotti,S., Nevanlinna,H., Butzow,R., Sobol, H., Jacquemier, J., Lyonet, D.S., Neuhausen, S.L., Weber, B., Wagner, T., Winqvist, R., Bignon, Y-J., Monti, F., Schmitt, F., Lenoir, G., Seitz, S., Hamman, U., Pharoah, P., Lane, G., Ponder, B., Bishop, D.T., and Easton, D.F. (2004) Pathology of Ovarian Cancers in *BRCA1* and *BRCA2* Carriers. Clin Cancer Res 10: **2473-2481**

Levy-Lahad, E. (2010) Fanconi anemia and breast cancer susceptibility meet again. Nat Genet. Vol. 42(5):**368-9**

Lewis, S. and Menon, U. (2004) Screening for ovarian cancer. Expert Rev Anticancer Ther. Vol. 3(1): **55-62**.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:**1754-60**.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics. Vol. 25 no. 16: **2078–2079**

Li, B. and Leal, S.M. (2009) Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies. Plos Genetics. Vol 5 (5) e1000481

Li, M., Reilly, M.P., Rader, D.J., and Wang, L-S. (2010) Correcting population stratification in genetic association studies using a phylogenetic approach. BIOINFORMATICS Vol. 26 no. 6, **798–806**

Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K. (2000) Environmental and heritable factors in the causation of cancer – analyses of cohorts of twins from Sweden, Denmark and Finland. N Engl J Med. Vol. 343;**78-85**

Liu, N., Lamerdin, J. E., Tebbs, R. S., Schild, D., Tucker, J. D., Shen, M. R., Brookman, K. W., Siciliano, M. J., Walter, C. A., Fan, W., Narayana, L. S., Zhou, Z.-Q., Adamson, A. W., Sorensen, K. J., Chen, D. J., Jones, N. J., Thompson, L. H. (1998) XRCC2 and XRCC3, new human Rad51-family members, promote chromosome stability and protect against DNA cross-links and other damages. Molec. Cell 1: **783-793**

Liu N, Schild D, Thelen MP, and Thompson LH. (2002) Involvement of Rad51C in two distinct protein complexes of Rad51 paralogs in human cells. Nucleic Acids Res. 2002 Vol. (4):1009-15.

Liu Y, Masson JY, Shah R, O'Regan P, and West SC. (2004) RAD51C is required for Holliday junction processing in mammalian cells. Science. Vol 9;303(5655):243-6.

Long, K.C., and Kauff, N.D. (2013) Screening for Familial Ovarian Cancer: A Ray of Hope and a Light to Steer by. JCO. Vol. 31 (1) **8-10**

Loveday, C., Turnbull, C., Ramsay, E., Hughes, D., Ruark, E., Frankum, J.R., Bowden, G., Kalmyrzaev, B., Warren-Perry, M., Snape, K., Adlard, J.W., Barwell, J., Berg, J., Brady, A.F., Brewer, C., Brice, G., Chapman, C., Cook, J., Davidson, R., Donaldson, A., Douglas, F., Greenhalgh, L., Henderson, A., Izatt, L., Kumar, A., Lalloo, F., Miedzybrodzka, Z., Morrison, P.J., Paterson, J., Mary Porteous, M., Rogers, M.T., Shanley, S., Walker, L., Breast Cancer Susceptibility Collaboration (UK), Eccles,D., Evans, D.G., Renwick, A., Seal, S., Lord, C.J., Ashworth, A., Reis-Filho, J.S., Antoniou, A.C., and Rahman, N. (2011) Germline mutations in *RAD51D* confer susceptibility to ovarian cancer. Nature Genetics. 43 (9):**879-882**

Lu, K. and Daniels, M (2013) Endometrial and ovarian cancer in women with Lynch Syndrome: update on screening and prevention Fam Cancer. (2):**273-7**

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, **1297-303**

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D.J. (2010) Target-enrichment strategies for next generation sequencing. Nat. Methods. Vol 7: **111-118**

Masson, J.-Y., Tarsounas, M. C., Stasiak, A. Z., Stasiak, A., Shah, R., McIlwraith, M. J., Benson, F. E., West, S. C. (2001) Identification and purification of two distinct complexes containing the five RAD51 paralogs. Genes Dev. 15: **3296-3307**

Mavaddat, N., Peock, S., Frost, D., Ellis, S., Platte, R., Fineberg, E., D. Evans, D.G., Izatt, L., A. Eeles, R.A., Adlard, J., Davidson. R., Eccles, D., Cole,T., Cook, J., Brewer, C., Tischkowitz, M., Douglas, F., Hodgson, S., Walker, L., Porteous, M.E Morrison, P.J., Side, L.E., Kennedy, M.J., Houghton, C., Donaldson, A., Rogers, M.T., Dorkins, H., Miedzybrodzka, Z., Gregory, H., Eason, J., Barwell, J., McCann, E., Murray, A., Antoniou, A.C., Easton, D.F., on behalf of EMBRACE (2013) Cancer Risks for BRCA1 and BRCA2 Mutation Carriers: Results From Prospective Analysis of EMBRACE JNCI J Natl Cancer Inst 105 (11): 812-822

Meindl A, Hellebrand H, Wiek C, Erven V, Wappenschmidt B, Niederacher D, Freund M, Lichtner P, Hartmann L, Schaal H, Ramser J, Honisch E, Kubisch C, Wichmann HE, Kast K, Deissler H, Engel C, Müller-Myhsok B, Neveling K, Kiechle M, Mathew CG, Schindler D, Schmutzler RK, Hanenberg H. (2010) Germline mutations in breast and ovarian cancer pedigrees establish *RAD51C* as a human cancer susceptibility gene. Nat Genet. 42(5): **410-4**

Menon, U., Gentry-Maharaj, A., Hallett, R., Ryan, A., Burnell, M., Sharma, A., Lewis, S., Davies, S., Philpott, S., Lopes, A., Godfrey, K., Oram, D., Herod, J., Williamson, K., Seif, M.W., Scott, I., Mould, T., Woolas, R., Murdoch, J., Dobbs, S., Amso, N.N., Leeson, S., Cruickshank, D., McGuire, A., Campbell, S., Fallowfield, L., Singh, N., Dawnay, A., Skates, S.J., Parmar, M. and Jacobs, I. (2009) Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Lancet Oncol. Vol. 10(4):**327-40**

Merritt, M.A., Green, A.C., Nagle, C.M., Webb, P.M., Australian Cancer Study(Ovarian Cancer) and Australian Ovarian Cancer Study Group (2008) Talcum powder, chronic pelvic inflammation and NSAIDs in relation to risk of epithelial ovarian cancer. Int. J. Cancer: 122, **170–176**

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu,Q., Cochran, C., Bennett, L.M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogden, R., Dayananth, P., Ward, J., Tonin, P., Narod, S., Bristow, P.K., Norris, F.H., Helvering, L., Morrison, P., Rosteck, P, Lai, M., Barrett, J.C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A., Skolnick, M.H. (1994) A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. Science, New Series, Vol. 266, No. 5182. 66-71

Milne, R.L., & Antoniou, A.C. (2011) Genetic modifiers of cancer risk for *BRCA1* and *BRCA2* mutation carriers Annals of Oncology 22 (Supplement 1): **i11–i17**

Miller, K. A., Sawicka, D., Barsky, D., and Albala, J.S. (2004) The domain mapping of the RAD51 protein complexes. Nucl. Acids Res. Vol. 32 (1): **169-178.**

Moody, A. (2007) An introduction to the polymerase chain reaction. In Hughes, S and Moody, A (Ed) (2007) PCR Methods Express. Scion Publishing.

Morgan, J.E., Carr, I.M., Sheridan, E., Chu, C.E., Hayward, B., Camm, N., Lindsay, H.A., Mattocks, C.J., Markham, A.F., Bonthron, D.T., and Taylor, G.R. (2010) Genetic diagnosis of familial breast cancer using clonal sequencing. Human Mutation. Vol. 31: No.4, 484-491

Mukhopadhyay, A., Elattar, A., Cerbinskaite,A., Wilkinson, S.J., Drew, Y, Kyle, S., Los, G., Hostomsky, Z., Edmonson, R.J., and Curtin, N.J., (2010) Development of a Functional Assay for Homologous Recombination Status in Primary Cultures of Epithelial Ovarian Tumor and Correlation with Sensitivity to Poly(ADP-Ribose) Polymerase Inhibitors Clin Cancer Res 16 (8) :**2344-2351**

Munroe, D.J., and Harris, T.J.R (2010) Third-generation sequencing fireworks at Marco Island. Nature Biotechnology 28, **426-428**

Nagy, R., Sweet, K., and Eng, C. (2004) Highly penetrant hereditary cancer syndromes Oncogene 23, **6445–6470**

Narod, S.A., and Foulkes, W.D., (2004) The BRCA1 network in response to DNA damage. Nat Review. Cancer (4): 665-676
Noralane M. Lindor, Mary L. McMaster, Carl J. Lindor, Mark H. Greene Concise Handbook of Familial Cancer Susceptibility Syndromes 2nd Edition (2008). Journal of the National Cancer Institute Monographs, No. 38

Ozcelik H, Shi X, Chang MC, Tram E, Vlasschaert M, Di Nicola N, Kiselova A, Yee D, Goldman A, Dowar M, Sukhu B, Kandel R, Siminovitch K. (2012) Long-Range PCR and Next-Generation Sequencing of BRCA1 and BRCA2 in Breast Cancer The Journal of Molecular Diagnostics Volume 14, Issue 5, **467-475**

Pal, T., Permuth-Wey, J., and Sellers, T.A. (2008) A Review of the Clinical Relevance of Mismatch-Repair Deficiency in Ovarian Cancer. Cancer. 113(4): **733–742**

Permuth-Wey, J., Lawrenson, K., Shen, H.C., Velkova, A., Tyrer JP, Chen Z, Lin HY, Chen YA, Tsai YY, Qu X, Ramus SJ, Karevan R, Lee J, Lee N, Larson MC, Aben KK, Anton-Culver H, Antonenkova N, Antoniou AC, Armasu SM; Australian Cancer Study; Australian Ovarian Cancer Study, Bacot F, Baglietto L, Bandera EV, Barnholtz-Sloan J, Beckmann MW, Birrer MJ, Bloom G, Bogdanova N, Brinton LA, Brooks-Wilson A, Brown R, Butzow R, Cai Q, Campbell I, Chang-Claude J, Chanock S, Chenevix-Trench G, Cheng JQ, Cicek MS, Coetzee GA; Consortium of Investigators of Modifiers of BRCA1/2, Cook LS, Couch FJ, Cramer DW, Cunningham JM, Dansonka-Mieszkowska A, Despierre E, Doherty JA, Dörk T, du Bois A, Dürst M, Easton DF, Eccles D, Edwards R, Ekici AB, Fasching PA, Fenstermacher DA, Flanagan JM, Garcia-Closas M, Gentry-Maharaj A, Giles GG, Glasspool RM, Gonzalez-Bosquet J, Goodman MT, Gore M, Górski B, Gronwald J, Hall P, Halle MK, Harter P, Heitz F, Hillemanns P, Hoatlin M, Høgdall CK, Høgdall E, Hosono S, Jakubowska A, Jensen A, Jim H, Kalli KR, Karlan BY, Kaye SB, Kelemen LE, Kiemeney LA, Kikkawa F, Konecny GE, Krakstad C, Kjaer SK, Kupryjanczyk J, Lambrechts D, Lambrechts S, Lancaster JM, Le ND, Leminen A, Levine DA, Liang D, Lim BK, Lin J, Lissowska J, Lu KH, Lubiński J, Lurie G, Massuger LF, Matsuo K, McGuire V, McLaughlin JR, Menon U, Modugno F, Moysich KB, Nakanishi T, Narod SA, Nedergaard L, Ness RB, Nevanlinna H, Nickels S, Noushmehr H, Odunsi K, Olson SH, Orlow I, Paul J, Pearce CL, Pejovic T, Pelttari LM, Pike MC, Poole EM, Raska P, Renner SP, Risch HA, Rodriguez-Rodriguez L,

Rossing MA, Rudolph A, Runnebaum IB, Rzepecka IK, Salvesen HB, Schwaab I, Severi G, Shridhar V, Shu XO, Shvetsov YB, Sieh W, Song H, Southey MC, Spiewankiewicz B, Stram D, Sutphen R, Teo SH, Terry KL, Tessier DC, Thompson PJ, Tworoger SS, van Altena AM, Vergote I, Vierkant RA, Vincent D, Vitonis AF, Wang-Gohrke S, Palmieri Weber R, Wentzensen N, Whittemore AS, Wik E, Wilkens LR, Winterhoff B, Woo YL, Wu AH, Xiang YB, Yang HP, Zheng W, Ziogas A, Zulkifli F, Phelan CM, Iversen E, Schildkraut JM, Berchuck A, Fridley BL, Goode EL, Pharoah PD, Monteiro AN, Sellers TA, Gayther SA. (2013) Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. NATURE COMMUNICATIONS. 4:1627

Pharoah, P.D., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F., and Ponder, B.A. (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. Vol. 31(1):**33-6**

Phelan CM, Kwan E, Jack E, Li S, Morgan C, Aubé J, Hanna D, and Narod SA. (2002) A low frequency of non-founder BRCA1 mutations in Ashkenazi Jewish breast-ovarian cancer families. Hum Mutat. Vol. 20(5):352-7

Piek, J.M.J., van Diest, P.J., Zweemer, R.P., Jansen, J.W., Poort-Keesom, R.J.J., Menko, F.H., Gille, J.J.P., Jongsma, A.P.M., Pals, G., Kenemans, P., Verheijen, R.H.M. (2001) Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer. The Journal of Pathology**.** 195(4); **451–456**

Piver, M.S, Mettlin, C.J, Tsukada, Y., Nasca, P., Greenwald, P., McPhee, M.E. (1984) Familial Ovarian Cancer Registry. Obstet Gynecol. 64(2):**195–199**.

Pray, L.A (2008) Embryo Screening and the Ethics of Human Genetic Engineering. Nature Education. Vol. 1 (1)

Press, J.Z., De Luca, A., Boyd, N., Young, S., Troussard, A., Ridge, Y., Kaurah, P., Kalloger, S.E., Blood, K.A., Smith, M., Spellman,P.T., Wang, Y., Miller, D.M., Horsman, D., Faham, M., Blake Gilks, C.B., Gray, J. and Huntsman, D.G.(2008) Ovarian carcinomas with genetic and epigenetic BRCA1 loss have distinct molecular abnormalities. BMC Cancer, 8:17

Rafnar**,** T**.**, Gudbjartsson. D.F**.**, Sulem, P., Jonasdottir, A., Sigurdsson, A., Jonasdottir,A., Besenbacher, S., Lundin, P., Stacey, S.N., Gudmundsson, J., Magnusson, O.T., le Roux, L., Orlygsdottir, G., Helgadottir, H.T., Johannsdottir, H., Gylfason, A., Tryggvadottir, L., Jonasson, J.G., de Juan, A., Ortega, E., Ramon-Cajal, J.M., García-Prats, M.D., Mayordomo, C., Panadero, A., Rivera, F., Aben, K,.K.H., van Altena**,** A.M., Massuger**,** L. F.A.G., Aavikko**,** M**.,** Kujala, P.M., Staff, S., Aaltonen, L.A., Olafsdottir, K**.,** Bjornsson, J**.,** Kong, A., Salvarsdottir, A., Saemundsson, H., Olafsson, K., Benediktsdottir, K.R**.,** Gulcher, J., Masson, G., Kiemeney, L.A., Mayordomo, J.I., Thorsteinsdottir, U. & Stefansson, K. (2011) Mutations in BRIP1 confer high risk of ovarian cancer. Nature Genetics. Vol 43 (11) 1104-1109

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., Jayatilake, H., McGuffog, L., Hanks, S., Evans, D. G., Eccles, D. (2007) The Breast Cancer Susceptibility Collaboration (UK), Easton, D. F., Stratton, M. R. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nature Genet. 39: **165-167**

Ratajska, M., Antoszewska, E., Piskorz, A., Brozek, I., Borg, Å., Kusmierek, H., Biernat, W., Limon, J. (2013) Cancer predisposing BARD1 mutations in breast-ovarian cancer families. Breast Cancer Res Treat. Vol. 131(1):89-97

Ramus SJ, Bobrow LG, Pharoah PD,, Finnigan, D.S., Fishman, A., Altaras, M., Harrington, P.A., Gayther, S.A., Ponder, B.A.J., Friedman, L.S. (1999) Increased frequency of *TP53* mutations in *BRCA1* and *BRCA2* ovarian tumours. Genes Chromosomes Cancer;25 (2):**91– 6**

Ramus, S.J. and Gayther, S.A. (2009) The Contribution of *BRCA1* and *BRCA2* to ovarian cancer. Molecular Oncology. Vol 3: **138-150**

Ramus, S.J., Harrington, P.A., Pye, C., DiCioccio, R.A., Cox, M.J., Garlinghouse-Jones, K., Oakley-Girvan, I., Jacobs, I.J., Hardy, R.M., Whittemore, A.S., Ponder, B. A. J., Piver, S., Pharoah, P.D.P, and Gayther. S.A. (2007) Contribution of *BRCA1* and *BRCA2* mutations to inherited ovarian cancer. Human Mutation. Vol 28(12): **1207-1215**

Rice, M. C., Smith, S. T., Bullrich, F., Havre, P., Kmiec, E. B. (1997) Isolation of human and mouse genes based on homology to REC2, a recombinational repair gene from the fungus Ustilago maydis. Proc. Nat. Acad. Sci. 94: **7417-7422**

Risch, H.A., McLaughlin, J.R., Cole, D.E., Rosen, B., Bradley, L., Fan, I., Tang, J., Li, S., Zhang, S., Shaw, P.A. and Narod, S.A. (2006) Population BRCA1 and BRCA2 mutation frequencies and cancer penetrances: a kin-cohort study in Ontario, Canada. J Natl Cancer Inst. Vol. 6;98(23):1694-706.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P (2011) Integrative genomics viewer. Nature Biotechnology. Vol 29 No.1 **24-26**

Roukos, D.H. (2009) Personalised cancer diagnostics and therapeutics. Expert Rev. Mol. Diagn. Vol. 9 (3) **227-229**

Sanger, F., Nicklen, S., and Coulson, A.R., (1977) Biochemistry DNA sequencing with chain-terminating inhibitors. PNAS. Vol. 74, No. 12, **5463-5467**

Schneegans, S.M., Rosenberger, A., Engel, U., Sander, M., Emons, G. and Shoukier, M. (2012) Validation of three BRCA1/2 mutation carrier probability models Myriad, BRCAPRO and BOADICEA in a population-based series of 183 German families. Familial Cancer. 11: **181-188**

Schoeman M, Apffelstaedt JP, Baatjes K, Urban M. (2013) Implementation of a breast cancer genetic service in South Africa - lessons learned. S Afr Med J. 103(8):**529-33**

Seidman JD, Horkayne-Szakaly I, Haiba M, Boice CR, Kurman RJ, Ronnett BM (2004) The histologic type and stage distribution of ovarian carcinomas of surface epithelial origin. Int J Gynecol Pathol. 23(1):41-4.

Smith, T.M., Lee, M.K., Szabo, C.I., Jerome, N., McEuen, M., Taylor, M., Hood, L., and King, M.C.(1996) Complete Genomic Sequence and Analysis of 117 kb of Human DNA Containing the Gene BRCA1. Cold Spring Harbor Laboratory Press GENOME RESEARCH Vol. 6:1029-1049

Soegaard, M., Kruger Kjaer, S., Cox, M., Wozniak, E., Hogdall, E., Hogdall, C., Blaakaer, J., Jacobs, I.J., Gayther, S.A., and Ramus, S.J. (2008) BRCA1 and BRCA2

Mutation Prevalence and Clinical Characteristics of a Population-Based Series of Ovarian Cancer Cases from Denmark. Human Cancer Biology. Clin Cancer Res. 14(12) 3761-3767

Speicher, M.R., Antonarakis, S.E., and Motulsky, A.G. (ed) (2010) Vogel and Motulsky's Human Genetics Problems and Approaches. 4th edition. Springer.
Stratton, M.R., & Rahman, N., (2008) The emerging landscape of breast cancer susceptibility. Nature Genetics. Vol 40 (1). 17-22

Strannenheim, H. and Lundeberg, J (2012) Stepping stones in DNA sequencing. Biotechnology Journal. Vol 7 (9): **1063-1073**

Stratton, J.F, Pharoah, P., Smith, S.K., Easton, D. and Ponder, B.A (1998) A systematic review and meta-analysis of family history and risk of ovarian cancer. Br J Obstet Gynaecol. Vol.105; **493-499**

Stratton, M.R. and Rahman, N (2008). The emerging landscape of breast cancer susceptibility. Nature genetics. Vol: 40, 1, pp17-22

Sulonen, A.M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A., and Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. Vol. 28;12(9):R94

Svendsen, J.M., Smogorzewska, A., Sowa, M.E., O'Connell, B. C., Gygi, S.P., Elledge, S.J., and Harper, J.W. (2009) Mammalian BTBD12/SLX4 Assembles A Holliday Junction Resolvase and Is Required for DNA Repair. Cell 138, **63–77**

Szabo, C., Masiello, A., Ryan, J.F., and Brody, L.C. (2000) The breast cancer information core: database design, structure, and scope. Hum Mutat. Vol.16(2):123-31

Tancredi, M., Sensi, E., Cipollini, G., Aretini, P., Lombardi, G., Di Cristofano, G.C., Presciuttini, S., Bevilacqua, G., and Caligo, M.A. (2004) Haplotype analysis of BRCA1 gene reveals a new gene rearrangement: characterization of a 19.9 KBP deletion. European Journal of Human Genetics. Vol 12, 775–777

Tavtigian, SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, Belanger C, Bell R, Berry S, Bogden R, Chen Q, Davis T, Dumont M, Frye C, Hattier T, Jammulapati S, Janecki T, Jiang P, Kehrer R, Leblanc JF, Mitchell JT, McArthur-Morrison J, Nguyen K, Peng Y, Samson C, Schroeder M, Snyder SC, Steele L, Stringfellow M, Stroup C, Swedlund B, Swense J, Teng D, Thomas A, Tran T, Tranchant M, Weaver-Feldhaus J, Wong AK, Shizuya H, Eyfjord JE, Cannon-Albright L, Tranchant M, Labrie F, Skolnick MH, Weber B, Kamb A, and Goldgar DE. (1996) The complete *BRCA2* gene and mutations in chromosome 13q-linked kindreds. Nature Genet. Vol. 12: **333-337**

The Cancer Genome Atlas (TCGA) Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature. Vol. 474, 609-615

Thompson D and Easton D; Breast Cancer Linkage Consortium (2001) Variation in cancer risks, by mutation position, in BRCA2 mutation carriers. Am J Hum Genet. Vol. (2):410-9.

Thompson ER, Doyle MA, Ryland GL, Rowley SM, Choong DY, Tothill RW, Thorne H; kConFab, Barnes DR, Li J, Ellul J, Philip GK, Antill YC, James PA, Trainer AH, Mitchell

G, and Campbell IG. (2012) Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. PLoS Genet. Vol.8 (9):e1002894

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2012) Integrative Genomic Viewer (IGV): high-performance genomics data visualisation and exploration. BRIEFINGS IN BIONIFOMRMATICS. Vol 14. No.2 **178-192**

Tischkowitz, M. and Xia, B. (2010) PALB2/FANCN: recombining cancer and Fanconi anemia. Cancer Res. Vol 1;70(19):**7353-9**

Tischkowitz, M., Capanu, M., Sabbaghian, N., Li, L., Liang, X., Vallee, M. P., Tavtigian, S. V., Concannon, P., Foulkes, W. D., Bernstein, L., The WECARE Study Collaborative Group, Bernstein, J. L., and Begg, C. B. (2012) Rare germline mutations in PALB2 and breast cancer risk: a population-based study. Hum. Mutat. Vol. 33: **674-680**

Tucker, T., Marra, M., and Friedman, J.M. (2009) Massively parallel sequencing: The next big thing in genetic medicine. Am J Hum Genetics. Vol 85 **142-154**

Venkitaraman, A.R., (2002) Cancer Susceptibility Review and the Functions of BRCA1 and BRCA2. Cell. Vol. 108, 171–182

Vogelstein, B. & Kinzler, K.W. (2004) Cancer genes and the pathways they control. Nat Med 10 (8): **789-99**

Walsh, T., Lee, M.K., Casadei, S., Thornton, A.M., Stray, S.M., Pennil, C., Nord, A.S., Mandell, J.B., Swisher, E.M., and King, M.C. (2010) Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. PNAS. Vol: 107, No.28, **12629-12633**

Walsh, T., Casadei, S., Lee, M.K., Pennil, C.C., Nord, A.S., Thornton, A.M., Roeb, W., Agnew, K.J., Stray, S.M., Wickramanayake, A., Norquist, B., Pennington, K.P., Garcia, R.L., King, M.C. and Swisher EM. (2011) Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proc Natl Acad Sci U S A. Vol. 1;108(44):**18032-7**

Wang Y, Cortez D, Yazdi P, Neff N, Elledge SJ, and Qin J. (2000) BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. Genes Dev. Vol. 14(8):927-39.

Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 2010, Vol. 38, No. 16

Weber W, Estoppey J, Stoll H. Familial cancer diagnosis. *Anticancer Res.* 2001;21(5):3631–3636.

Whittemore, A.S, Gong G, John EM, McGuire V, Li FP, Ostrow KL, Dicioccio R, Felberg A, West DW (2004) Prevalence of BRCA1 Mutation Carriers among U.S. Non-Hispanic Whites. Cancer Epidemiol Biomarkers Prev.13(12)

Wilkinson, E. (2012) Preimplantation genetic diagnosis for mutated BRCA genes. Lancet Oncology. Vol. 13 e331

Wilson JB, Yamamoto K, Marriott AS, Hussain S, Sung P, Hoatlin ME, Mathew CG, Takata M, Thompson LH, Kupfer GM, and Jones NJ.(2008) FANCG promotes formation of a newly identified protein complex containing BRCA2, FANCD2 and XRCC3. Oncogene. Vol.12; 27(26):**3641-52**

Wu, L. C., Wang, Z. W., Tsan, J. T., Spillman, M. A., Phung, A., Xu, X. L., Yang, M.-C. W., Hwang, L.-Y., Bowcock, A. M., and Baer, R. (1996) Identification of a RING protein that can interact in vivo with the BRCA1 gene product. Nature Genet. 14: 430-440

Yang, D. Khan, S. Sun, Y., Hess, K., Shmulevich, I., Sood, A.K., Zhang, W. (2011) Association of BRCA1 and BRCA2 Mutations With Survival, Chemotherapy Sensitivity, and Gene Mutator Phenotype in Patients With Ovarian Cancer JAMA. 306(14):**1557-1565**

Yoon, S., Xuan,Z., Makarov, V., Ye, K., and Sebat, J., (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research. 19:1586–1592

Zhong, Q., Chen, C.-F., Li, S., Chen, Y., Wang, C.-C., Xiao, J., Chen, P.-L., Sharp, Z. D., and Lee, W.-H. (1999) Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. Science 285: 747-750,.

http://biowulf.nih.gov/apps/GATK.html

http://info.cancerresearchuk.org/cancerstats/types/ovary/survival/#stage accessed 28-09-10
http://www.nice.org.uk/nicemedia/pdf/CG41NICEguidance.pdf accessed 28-09-10

http://www.nice.org.uk/nicemedia/pdf/CG41NICEguidance.pdf accessed 28-09-10
http://info.cancerresearchuk.org/cancerstats/types/ovary/survival/#stage accessed 28-09-10

http://bcb.dfci.harvard.edu/bayesmendel/brcapro.php accessed 11-12-12
CASAVA v1.6 User Guide:
http://atlas.bx.psu.edu/_static/pdf/CASAVA1.6_User_Guide_15009919_A.pdf

http://watson.nci.nih.gov/solexa/CASAVA1.6_User_Guide_15009919_A.pdf accessed 24_01-11

http://hapmap.ncbi.nlm.nih.gov/ accessed 05_01_11

www.hfea.gov.uk (Human Fertility and Embryology Authority) accessed 05-09-2013

http://research.nh- gri.nih.gov/bic/

http://www.illumina.com/Documents/%5Cproducts%5Ctechnotes%5Ctechnote_Q-Scores.pdf accessed 16-04-13

http://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf accessed 16-04-13

http://www.instituteforwomenshealth.ucl.ac.uk/academic_research/gynaecologicalcancer/gcrc/ukops/health accessed 02-08-2013

Polish Ovarian cancer case control study http://dceg.cancer.gov/research/cancer-types/ovary/ovarian-endometrial-cancer-case-control-study-poland accessed 02-08-2013

Genome analysis toolkit http://www.broadinstitute.org/gatk/index.php

http://www.instituteforwomenshealth.ucl.ac.uk/academic_research/gynaecologicalcancer/gcrc/ukfocss/eligibility accessed 09_08_2013

European Science Foundation document (2012) 'Personalised Medicine for the European Citizen – Towards more precise medicine for the diagnosis, treatment and prevention of disease (iPM)' sourced from:
http://www.esf.org/uploads/media/Personalised_Medicine.pdf accessed 05-09-2013

UPMC Cancer Centre http://www.upmccancercenter.com/genetic/guide.cfm accessed 05-09-2013

http://www.ambrygen.com/tests/ovanext 08-09-2013

http://www.ncbi.nlm.nih.gov/tools/primer-blast/
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
http://www.broadinstitute.org/gatk/
http://www.broadinstitute.org/igv/
http://www. http://provean.jcvi.org
http://www.illumina.com/systems/hiseq_2000_1000.ilmn

http://bcb.dfci.harvard.edu/BayesMendel/ accessed 09-09-2012

Appendix I

Table of LR-PCR primer sequences and their position on the reference sequence L78833.1

Fragment	Primer	Sequence	Start	Stop	Size Kb
BRCA1_1 Prom-intron 2	Forward	AGTGTCACCACCCCAAGGACTCTCT	122	146	9.5
	Reverse	ACGACTAACCTGGCAGTGTGACAAG	9623	9599	
BRCA1_2 Intron 2-3	Forward	ACTGTCCACAAGCTTTTCTTGTGATCC	9415	9441	9.5
	Reverse	ACAATTCAGAGCAGGGGTAGGGAGG	19000	18976	
BRCA1_3BFint3 BRCA1_3BRint6	Forward	TGTATAGACTACAGCACGAGACAGCTT	18898	18924	5.009
	Reverse	ACAGCACTTGAGTTGCATTCTTGGG	23906	23882	
BRCA1_3CFint6 BRCA1_3DRint8	Forward	TGTGCTTTTCAGCTTGACACAGGT	23836	23859	5.177
	Reverse	AGCTCTTCTTAAAAGGCTTCCTCATCTAGT	29606	29577	
BRCA1_4A Intron 7-Exon 11	Forward	CCAGACATTTTAGTGTGTAAATTCCTGGGC	28686	28715	8.415
	Reverse	CACTGTGAAGAAAACAAGCTAGCAGAAC	37100	37073	
BRCA1_11BF BRCA1_13R	Forward	AGCCCTTTCACCCATACACA	36772	36791	9.411
	Reverse	CAGGTTATGTTGCATGGTATC	46162	46183	
BRCA1_5A Intron 12- 13	Forward	ACATCAAGTCTATTTGGGGGAATTTGAGG	45915	45943	6.047
	Reverse	TGCAACAGACAGATGCTAGCACCAAA	51961	51936	
BRCA1_6A Intron 13-16	Forward	AGCCTTGTCTCAGCTGGGTGTCT	51849	51871	6.436
	Reverse	GCAGGGCAGAAGTGGCAGGG	58284	58265	
BRCA1_7 Intron 16-19	Forward	GGCTTGTAAGAATGCCCTGCCACTT	58251	58275	9.2
	Reverse	CCTAGTGCCCAGAACACAGTAGGCT	67271	67247	
BRCA1_8 Intron 19-20	Forward	ATTGGGAGCCTACTGTGTTCTGGG	67241	67264	10.0
	Reverse	AGAGGCTTGGATGGCTAGAAACTCA	77242	77218	
BRCA1_9 Intron 20 to beyond polyA tail	Forward	AGAGGAGACAAGGAGCATGTACACC	77102	77126	10.0
	Reverse	CAGGTTGCTGGCCCCACCTGTCTGG	87087	87063	

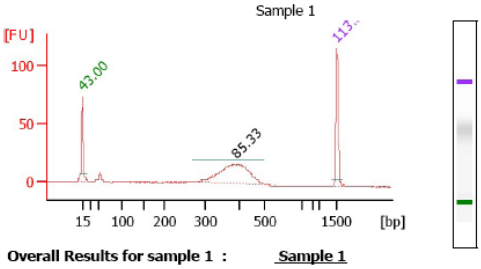
Appendix II

Agilent trace output of 12 samples

Assay Class: DNA 1000
Data Path: Y:\...8\JH_E+C_180510_DNA 1000_DE24802286_2010-05-18_09-18-36.xad
Created: 5/18/2010 9:18:36 A
Modified: 5/19/2010 2:21:47 F

Electropherogram Summary

Sample 1



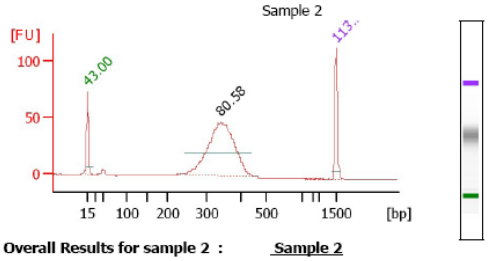
Overall Results for sample 1 : Sample 1

Number of peaks found: 1

Peak table for sample 1 : Sample 1

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	390	4.76	18.5	
3	1,500	2.10	2.1	Upper Marker

Sample 2



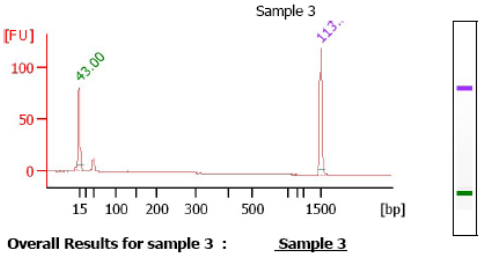
Overall Results for sample 2 : Sample 2

Number of peaks found: 1

Peak table for sample 2 : Sample 2

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	341	13.53	60.1	
3	1,500	2.10	2.1	Upper Marker

Sample 3



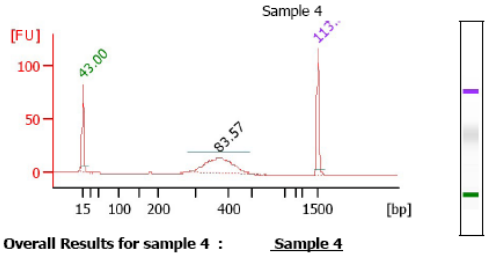
Overall Results for sample 3 : Sample 3

Number of peaks found: 0

Peak table for sample 3 : Sample 3

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	1,500	2.10	2.1	Upper Marker

Sample 4



Overall Results for sample 4 : Sample 4

Number of peaks found: 1

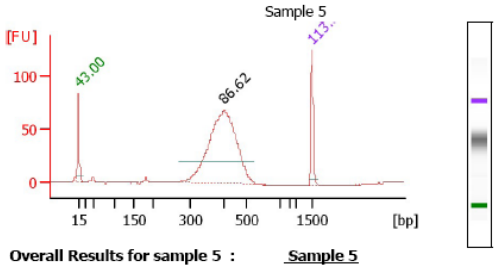
Peak table for sample 4 : Sample 4

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	372	4.03	16.4	
3	1,500	2.10	2.1	Upper Marker

Assay Class: DNA 1000
Data Path: Y:\...8\JH_E+C_180510_DNA 1000_DE24802286_2010-05-18_09-18-36.xad

Created: 5/18/2010 9:18:36 AM
Modified: 5/19/2010 2:21:47 PM

Electropherogram Summary Continued ...

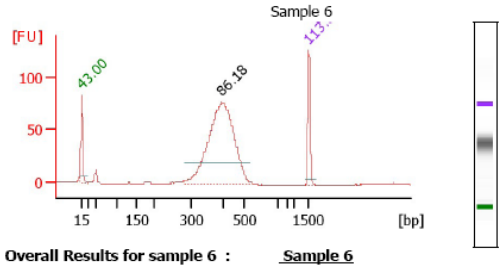


Overall Results for sample 5 : Sample 5

Number of peaks found: 1

Peak table for sample 5 : Sample 5

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	404	17.21	64.5	
3	1,500	2.10	2.1	Upper Marker

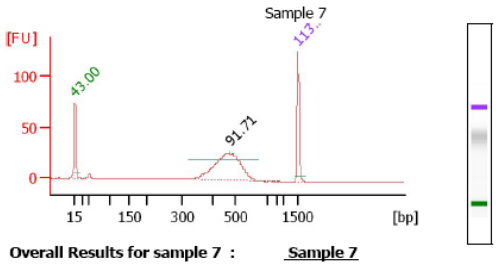


Overall Results for sample 6 : Sample 6

Number of peaks found: 1

Peak table for sample 6 : Sample 6

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	399	18.13	68.9	
3	1,500	2.10	2.1	Upper Marker

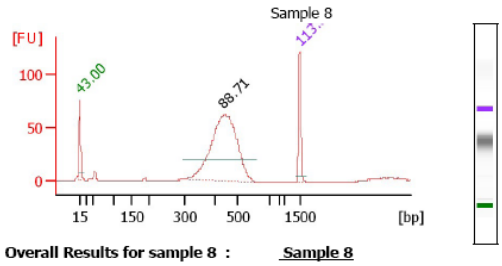


Overall Results for sample 7 : Sample 7

Number of peaks found: 1

Peak table for sample 7 : Sample 7

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	478	6.23	19.7	
3	1,500	2.10	2.1	Upper Marker



Overall Results for sample 8 : Sample 8

Number of peaks found: 1

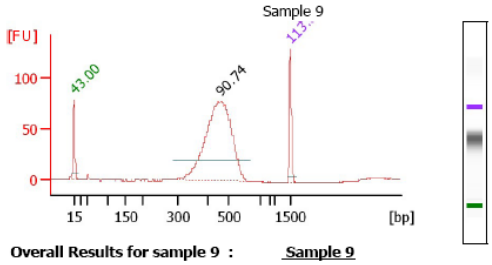
Peak table for sample 8 : Sample 8

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	435	15.29	53.3	
3	1,500	2.10	2.1	Upper Marker

Assay Class: DNA 1000
Data Path: Y:\...8\JH_E+C_180510_DNA 1000_DE24802286_2010-05-18_09-18-36.xad

Created: 5/18/2010 9:18:36 AM
Modified: 5/19/2010 2:21:47 PM

Electropherogram Summary Continued ...

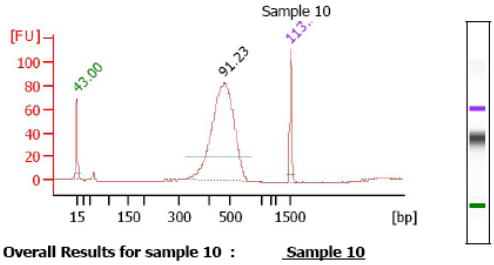


Overall Results for sample 9 : **Sample 9**

Number of peaks found: 1

Peak table for sample 9 : **Sample 9**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	464	18.24	59.6	
3	1,500	2.10	2.1	Upper Marker

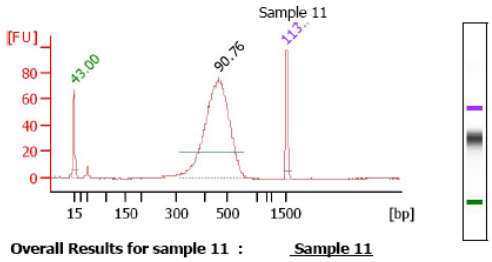


Overall Results for sample 10 : **Sample 10**

Number of peaks found: 1

Peak table for sample 10 : **Sample 10**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	471	20.31	65.3	
3	1,500	2.10	2.1	Upper Marker

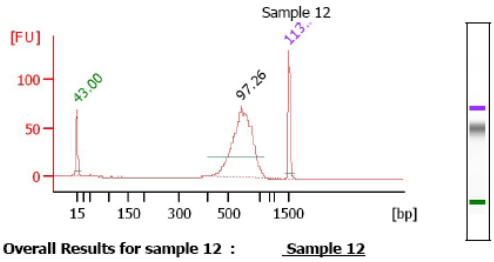


Overall Results for sample 11 : **Sample 11**

Number of peaks found: 1

Peak table for sample 11 : **Sample 11**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	464	23.33	76.1	
3	1,500	2.10	2.1	Upper Marker



Overall Results for sample 12 : **Sample 12**

Number of peaks found: 1

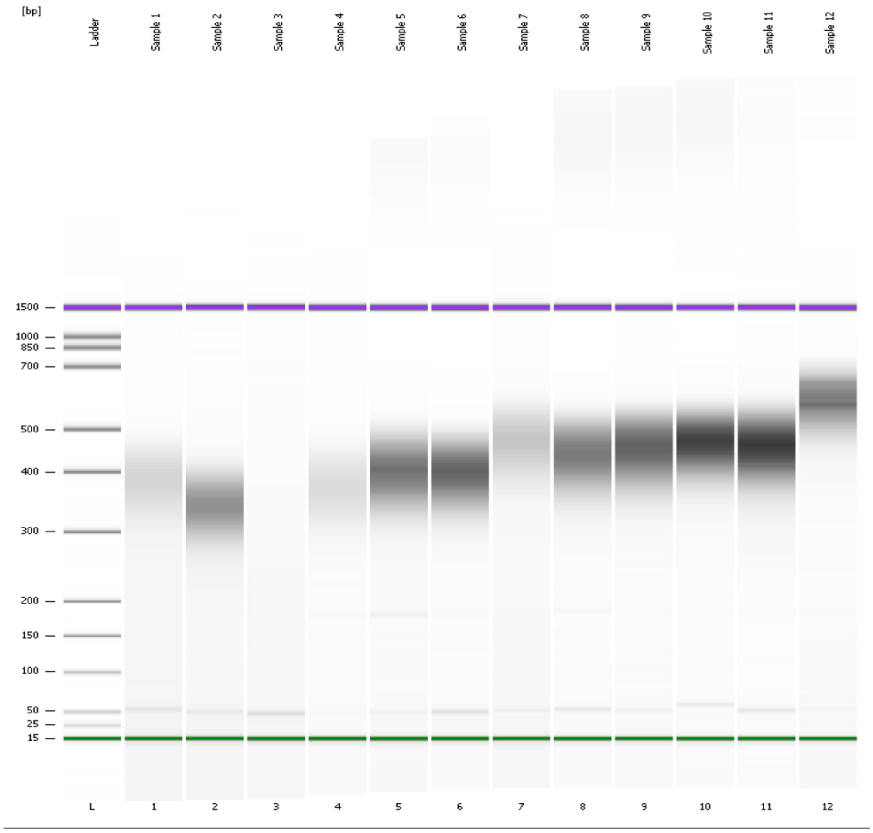
Peak table for sample 12 : **Sample 12**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	580	11.87	31.0	
3	1,500	2.10	2.1	Upper Marker

Assay Class: DNA 1000
Data Path: Y:\...8\JH_E+C_180510_DNA 1000_DE24802286_2010-05-18_09-18-36.xad

Created: 5/18/2010 9:18:36 AM
Modified: 5/19/2010 2:21:47 PM

Gel Image

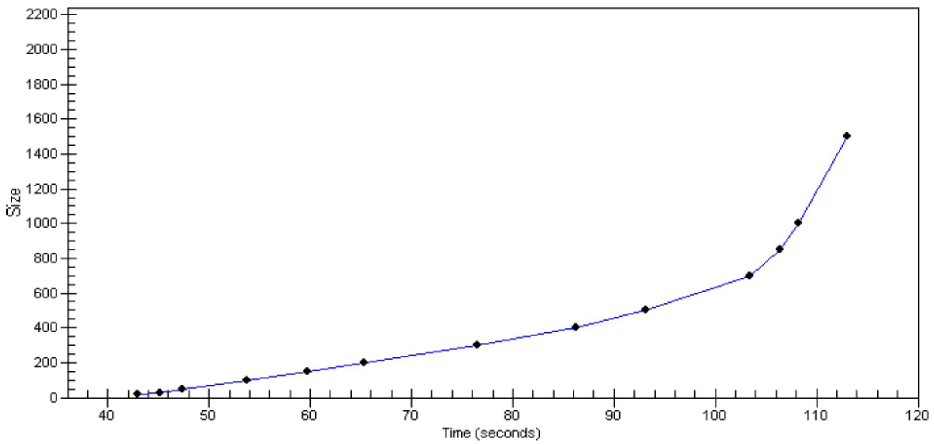


Assay Class: DNA 1000
Data Path: Y:\...8\JH_E+C_180510_DNA 1000_DE24802286_2010-05-18_09-18-36.xad

Created: 5/18/2010 9:18:36 AM
Modified: 5/19/2010 2:21:47 PM

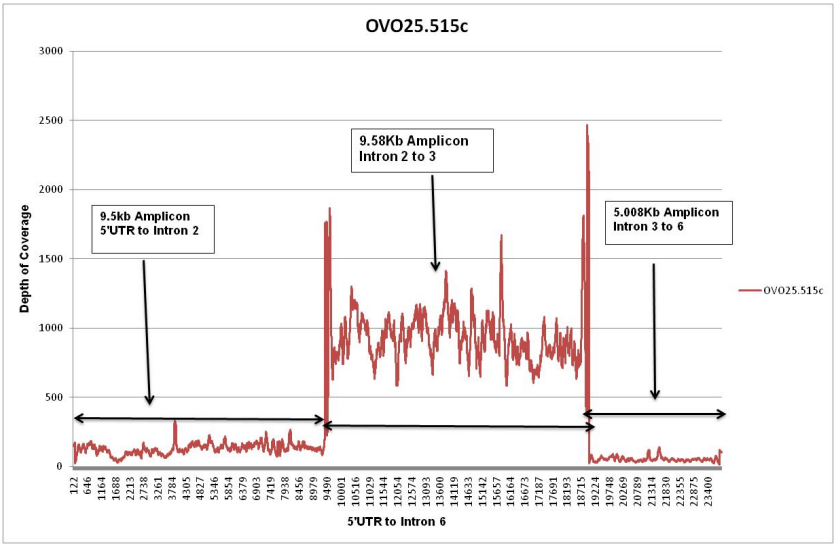
Curves

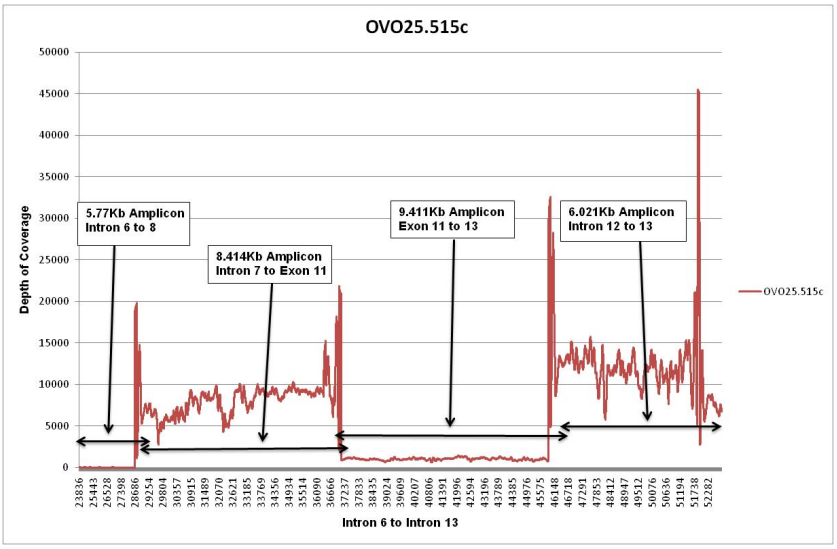
Standard Curve

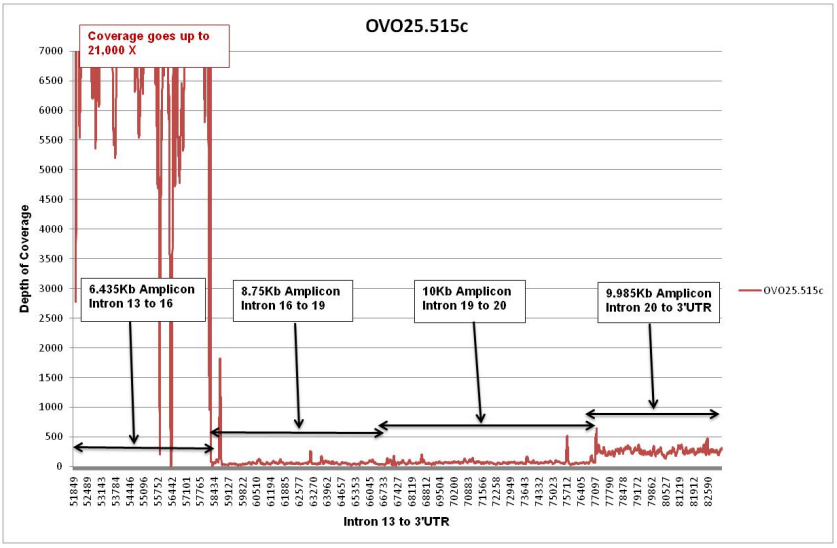


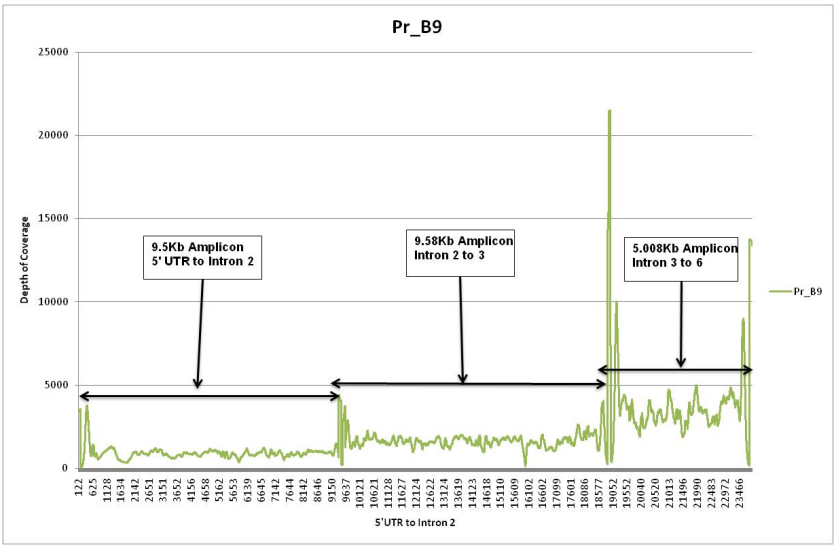
Appendix III

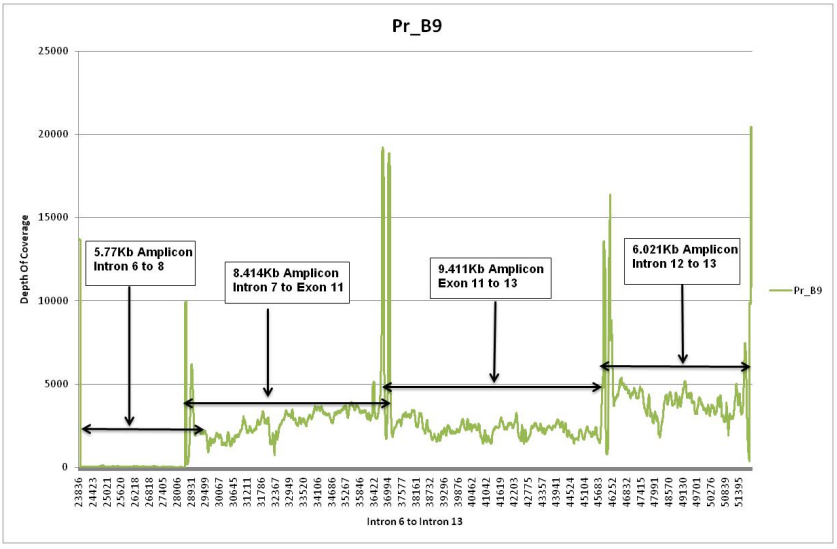
Coverage Data for 11 Samples

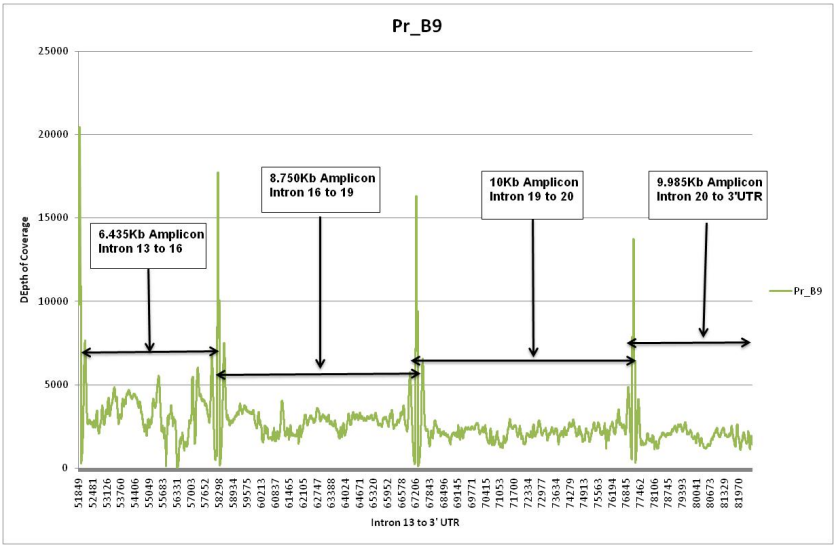


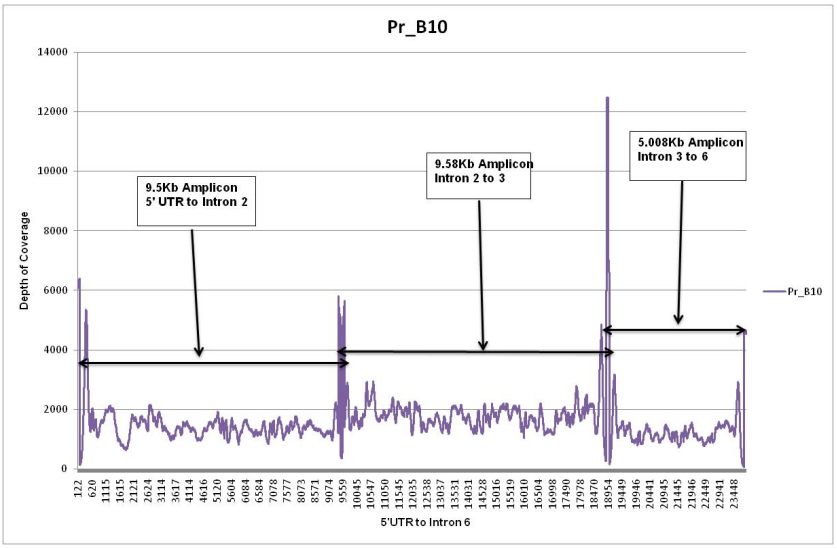


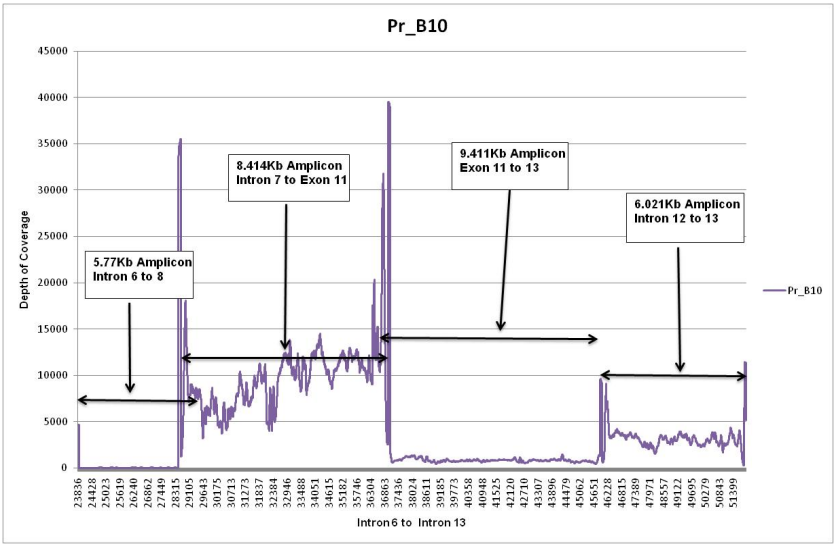


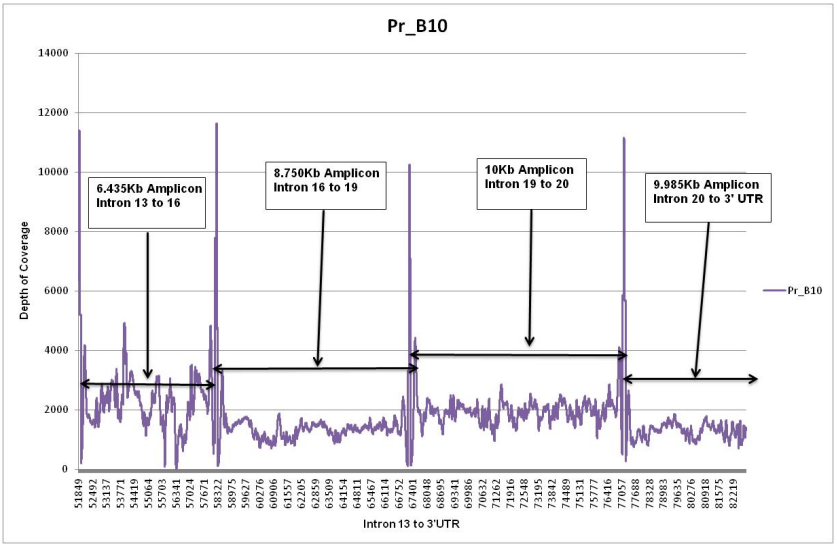

Pr_B9
Depth of Coverage
9.5Kb Amplicon
5' UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Pr_B9
5'UTR to Intron 2

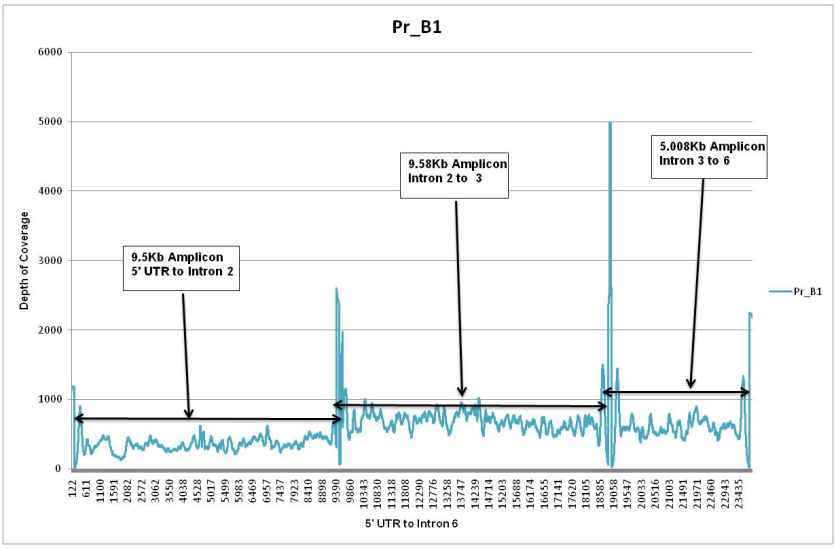

Pr_B9
Depth Of Coverage
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Pr_B9
Intron 6 to Intron 13

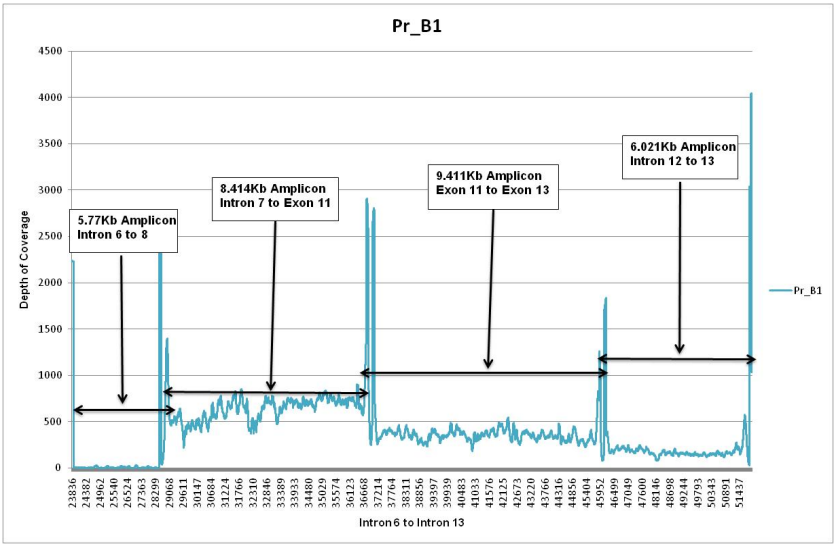

Pr_B9
Depth of Coverage
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Pr_B9
Intron 13 to 3' UTR

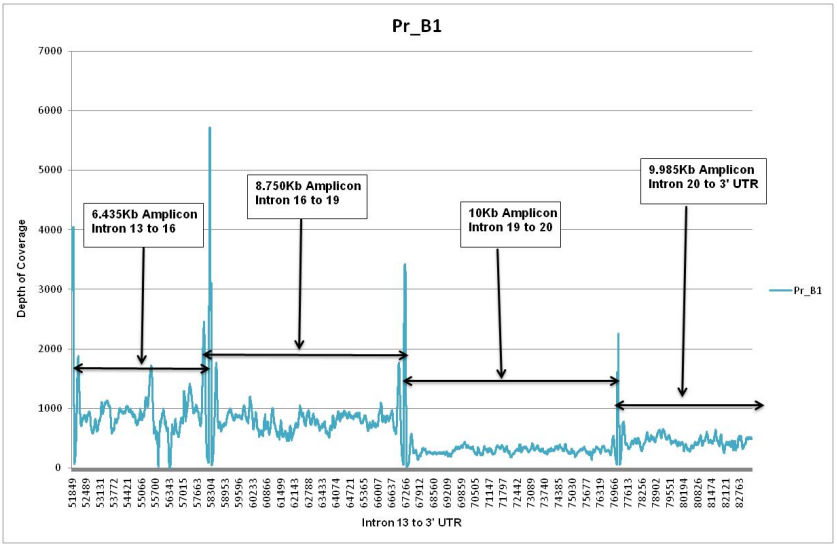

Pr_B10
Depth of Coverage
9.5Kb Amplicon
5' UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Pr_B10
5'UTR to Intron 6

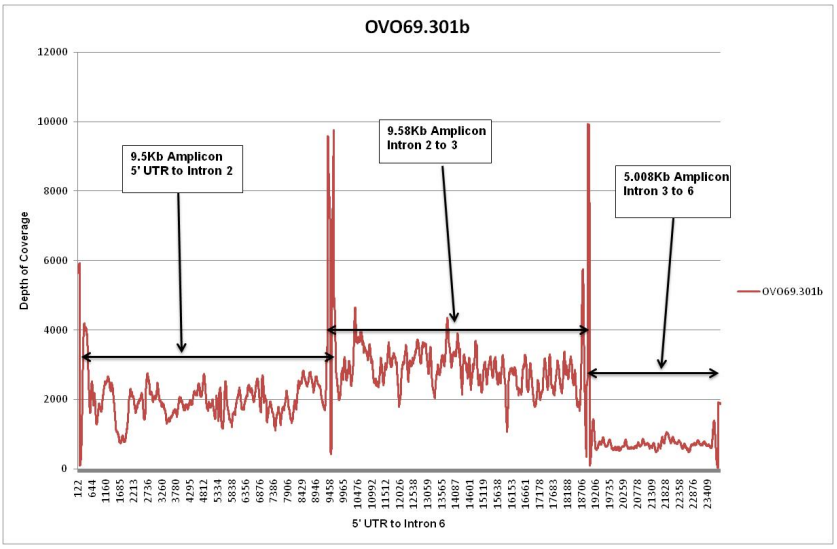

Pr_B10
Depth of Coverage
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Pr_B10
Intron 6 to Intron 13

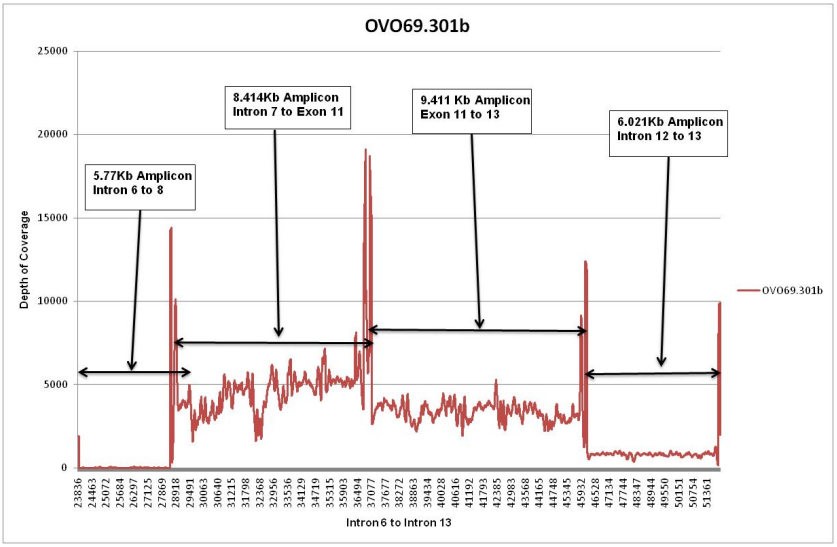

Pr_B10
Depth of Coverage
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3' UTR
Pr_B10
Intron 13 to 3'UTR

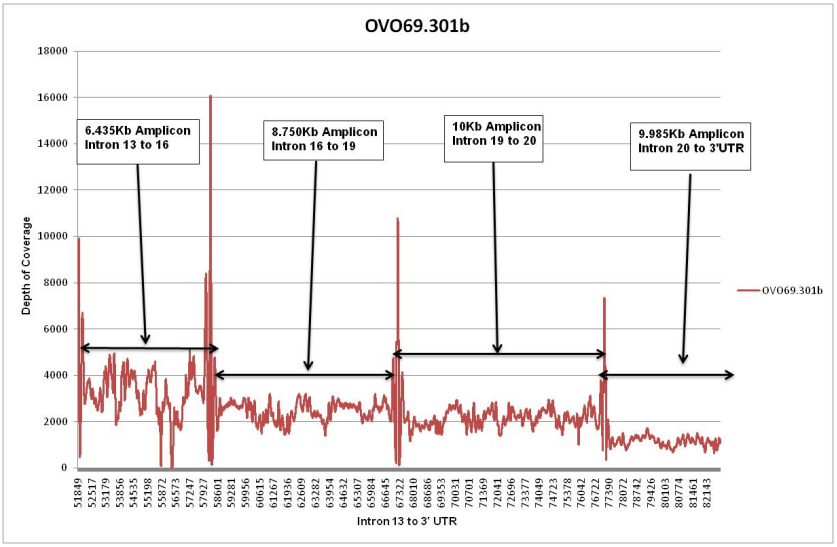

Pr_B1
Depth of Coverage
9.5Kb Amplicon
5' UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Pr_B1
5' UTR to Intron 6

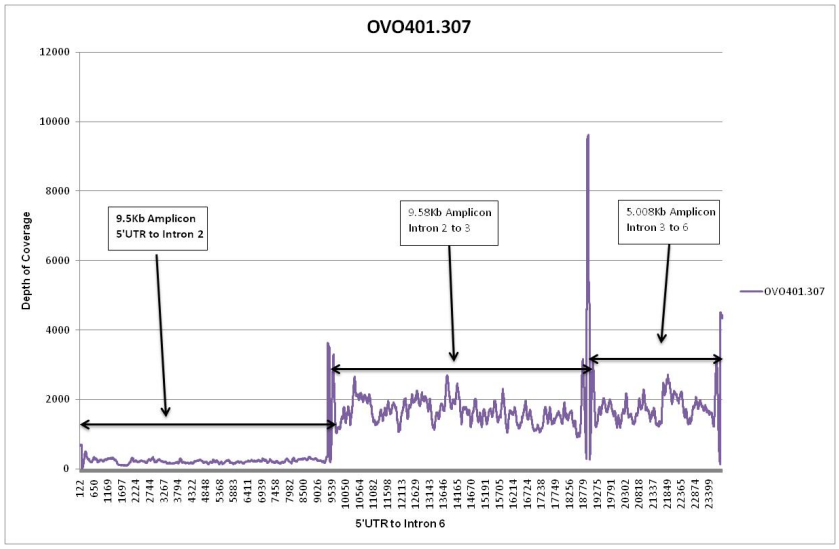

Pr_B1
Depth of Coverage
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to Exon 13
6.021Kb Amplicon
Intron 12 to 13
Pr_B1
Intron 6 to Intron 13

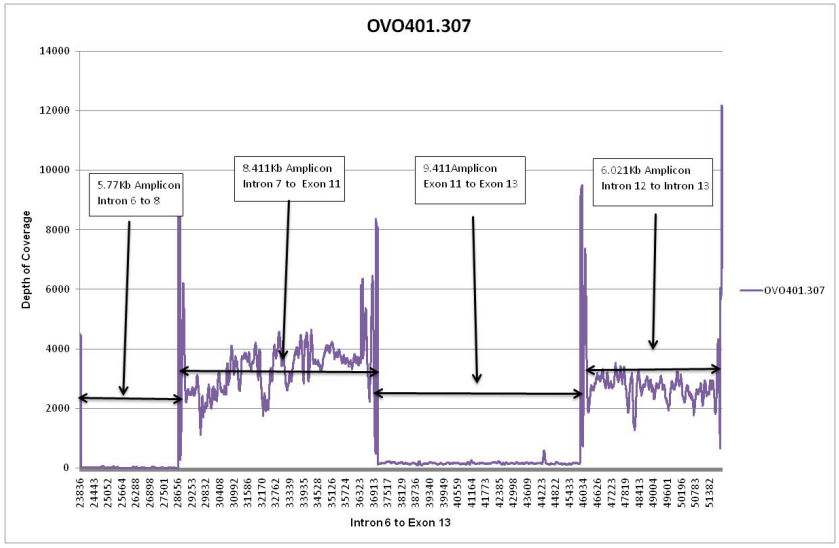

Pr_B1
Depth of Coverage
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3' UTR
Pr_B1
Intron 13 to 3' UTR

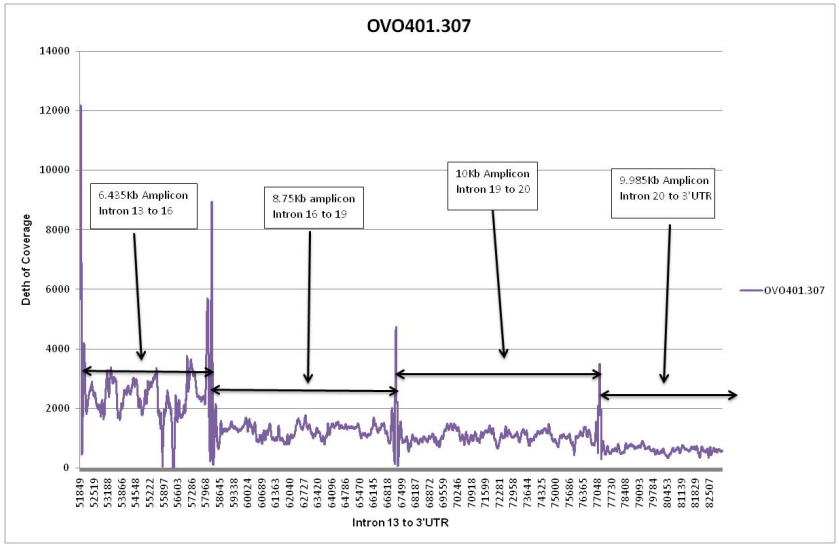

OVO69.301b
9.5Kb Amplicon
5' UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Depth of Coverage
OVO69.301b
5' UTR to Intron 6

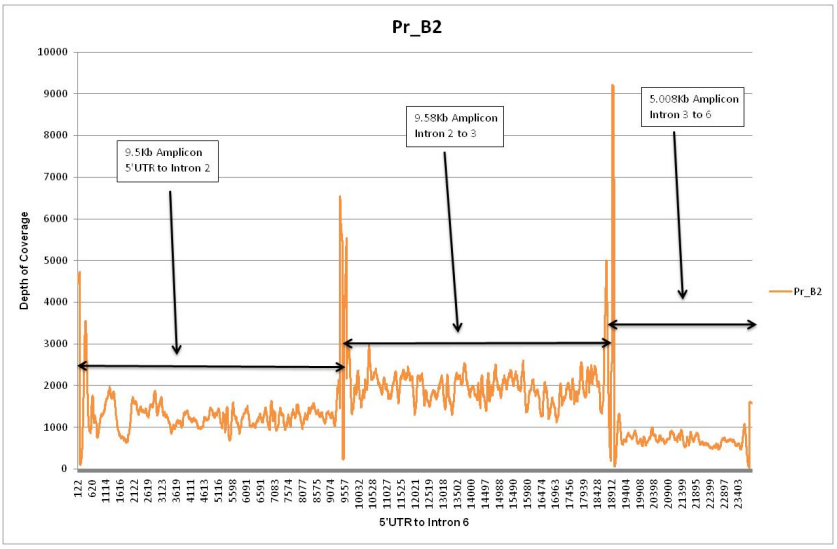

OVO69.301b
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411 Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Depth of Coverage
OVO69.301b
Intron 6 to Intron 13

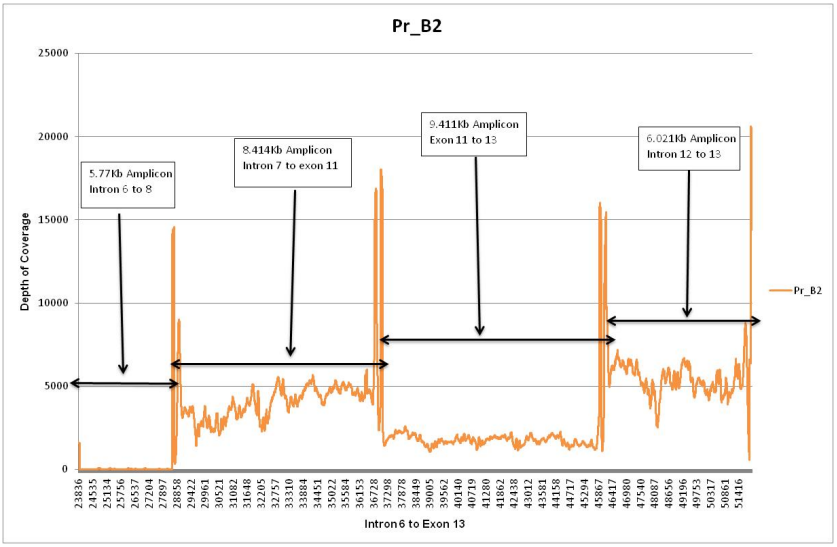

OVO69.301b
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Depth of Coverage
OVO69.301b
Intron 13 to 3' UTR

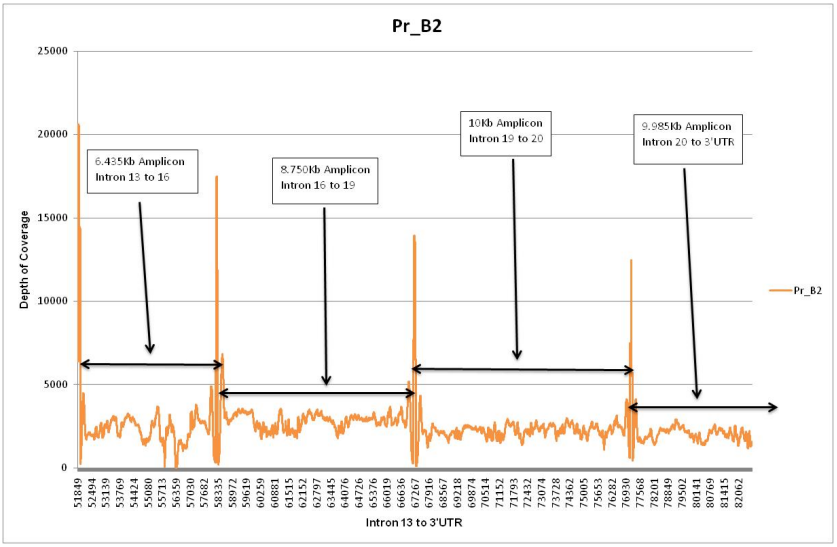

OVO401.307
Depth of Coverage
9.5Kb Amplicon
5'UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
OVO401.307
5'UTR to Intron 6

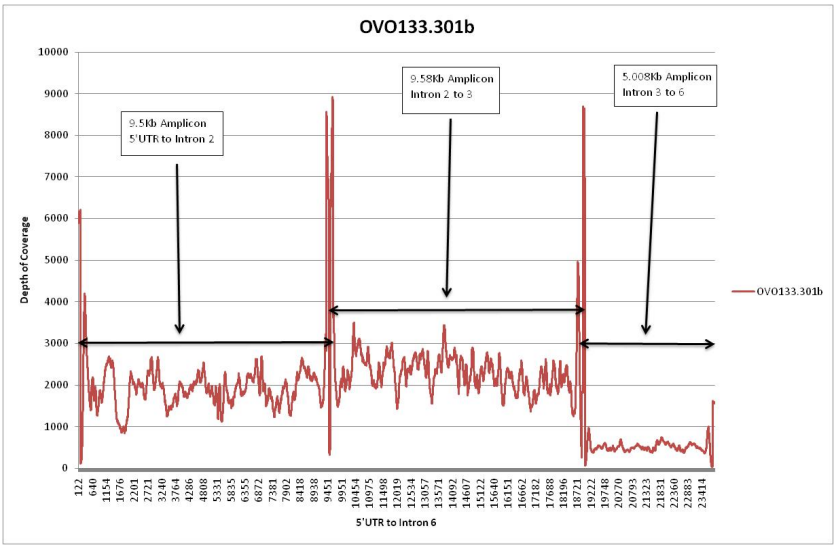

OVO401.307
Depth of Coverage
5.77Kb Amplicon
Intron 6 to 8
8.411Kb Amplicon
Intron 7 to Exon 11
9.411Amplicon
Exon 11 to Exon 13
6.021Kb Amplicon
Intron 12 to Intron 13
OVO401.307
Intron 6 to Exon 13

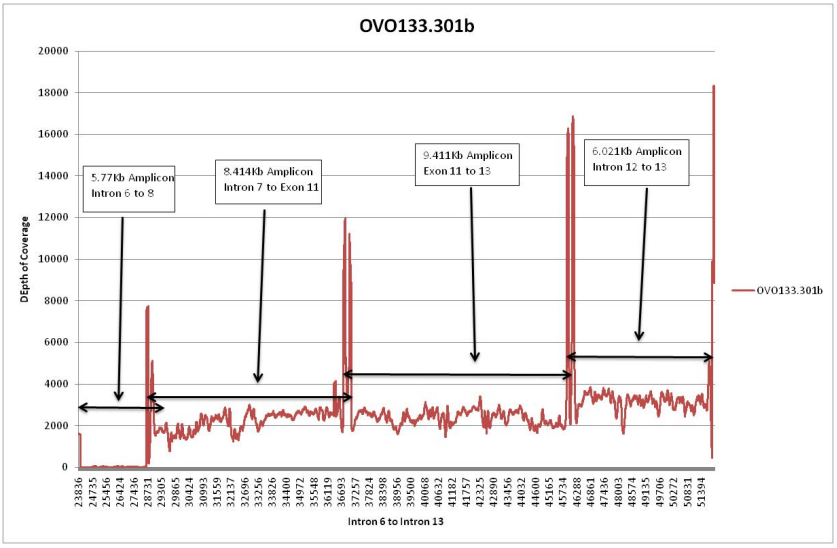

OVO401.307
Depth of Coverage
6.435Kb Amplicon
Intron 13 to 16
8.75Kb amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
OVO401.307
Intron 13 to 3'UTR

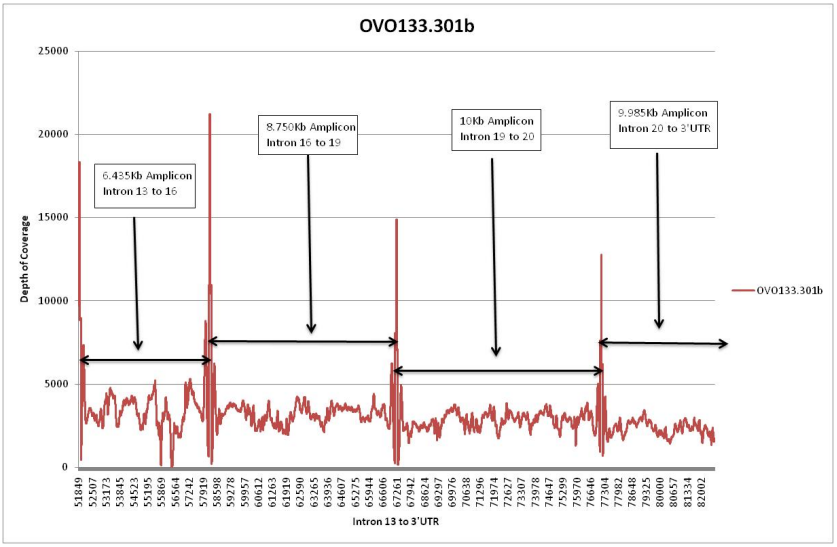

Pr_B2
9.5Kb Amplicon
5'UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Depth of Coverage
Pr_B2
5'UTR to Intron 6

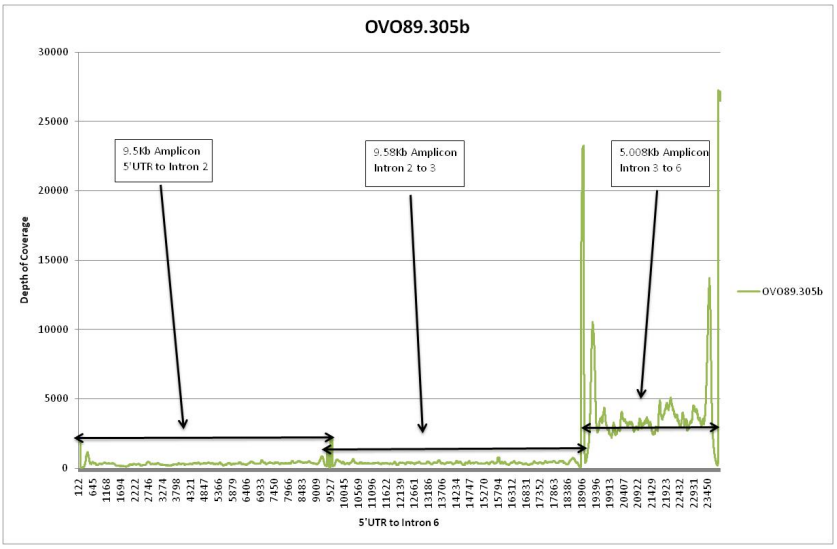

Pr_B2
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Depth of Coverage
Pr_B2
Intron 6 to Exon 13

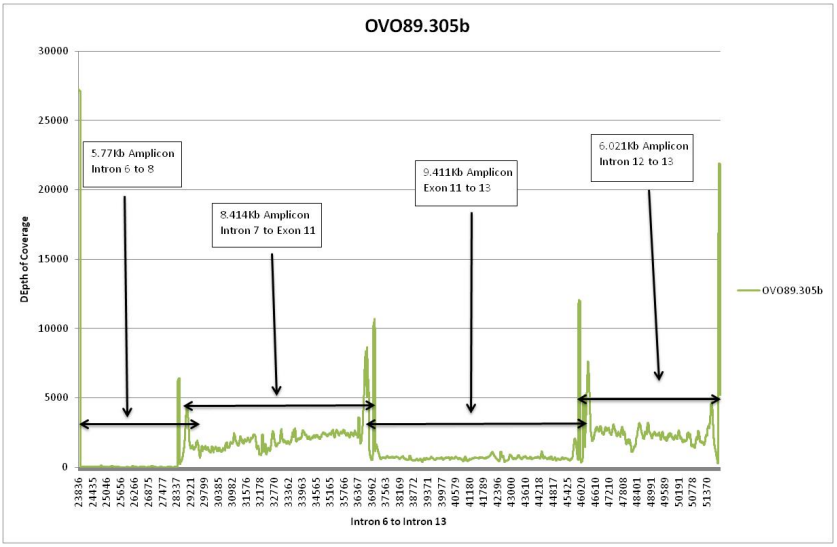

Pr_B2
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Depth of Coverage
Pr_B2
Intron 13 to 3'UTR

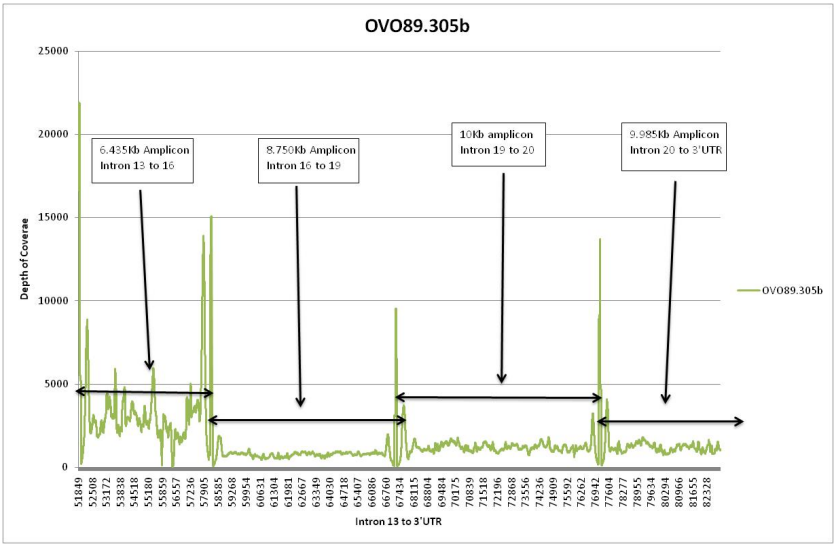

OVO133.301b
9.5Kb Amplicon
5'UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Depth of Coverage
OVO133.301b
5'UTR to Intron 6

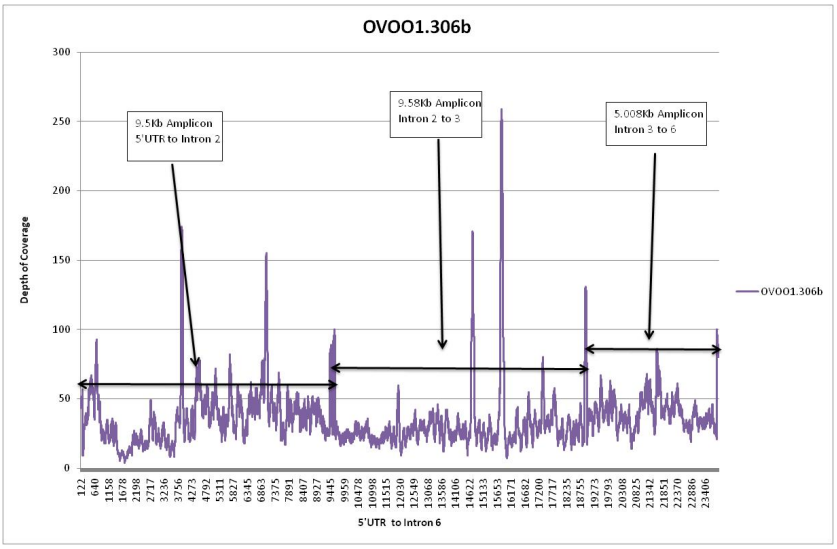

OVO133.301b
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Depth of Coverage
OVO133.301b
Intron 6 to Intron 13

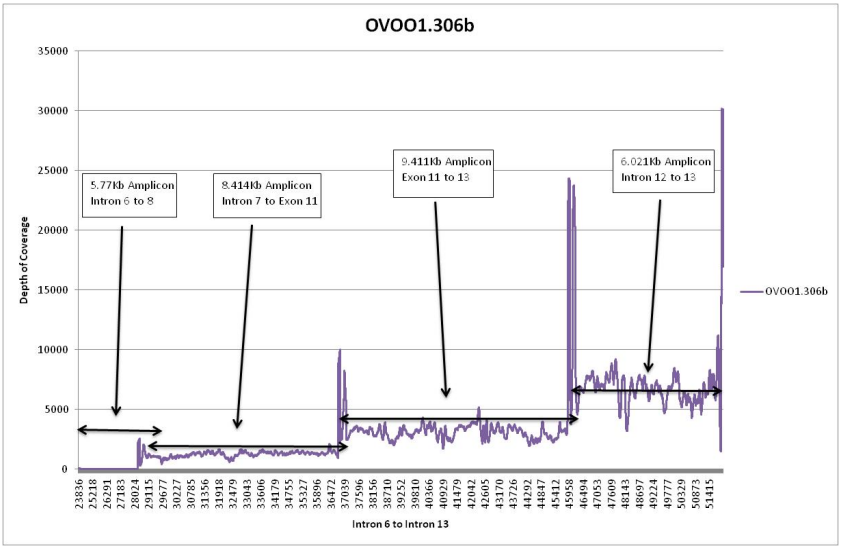

OVO133.301b
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Depth of Coverage
OVO133.301b
Intron 13 to 3'UTR

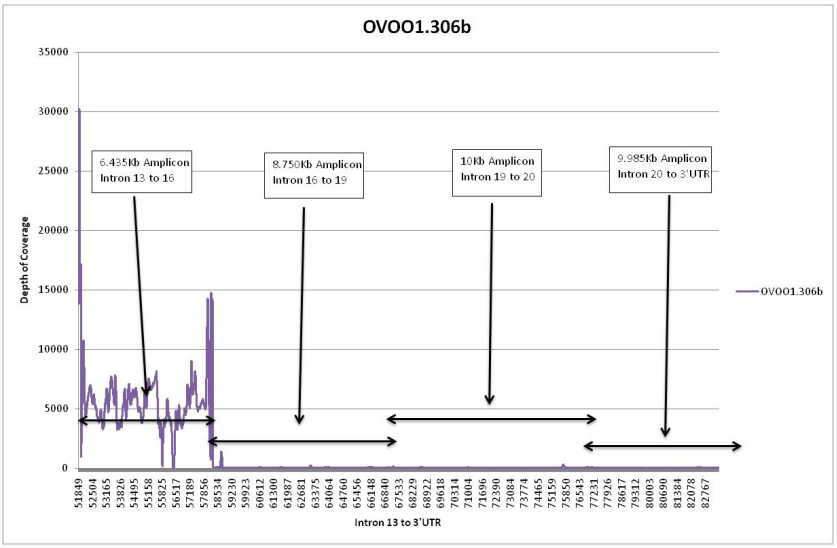

OVO89.305b
9.5Kb Amplicon
5'UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Depth of Coverage
OVO89.305b
5'UTR to Intron 6

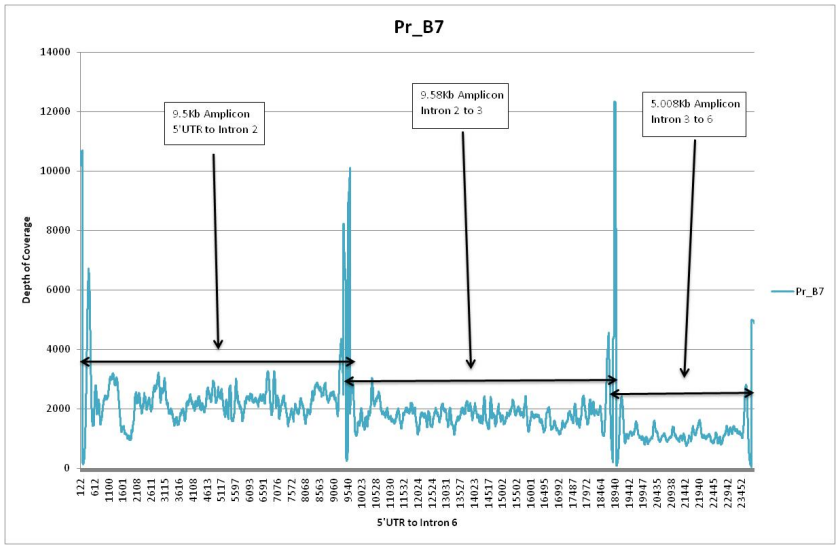

OVO89.305b
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Depth of Coverage
OVO89.305b
Intron 6 to Intron 13

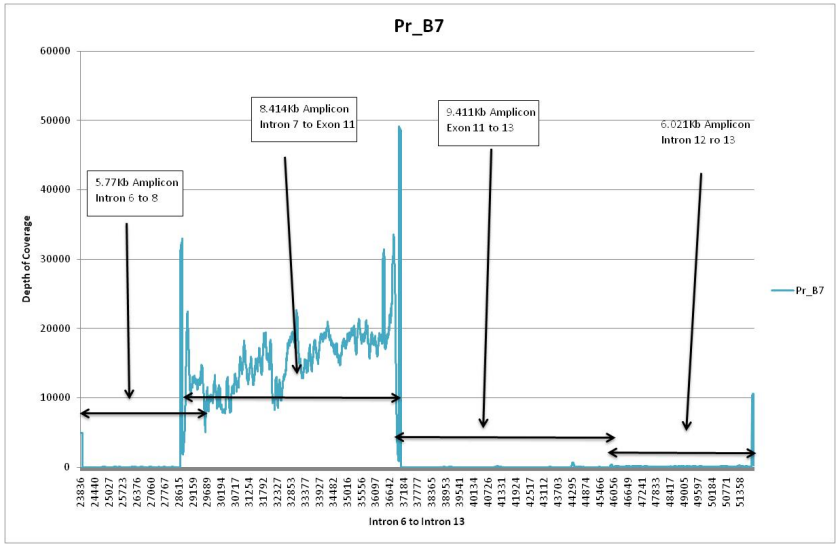

OVO89.305b
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Depth of Coverage
OVO89.305b
Intron 13 to 3'UTR

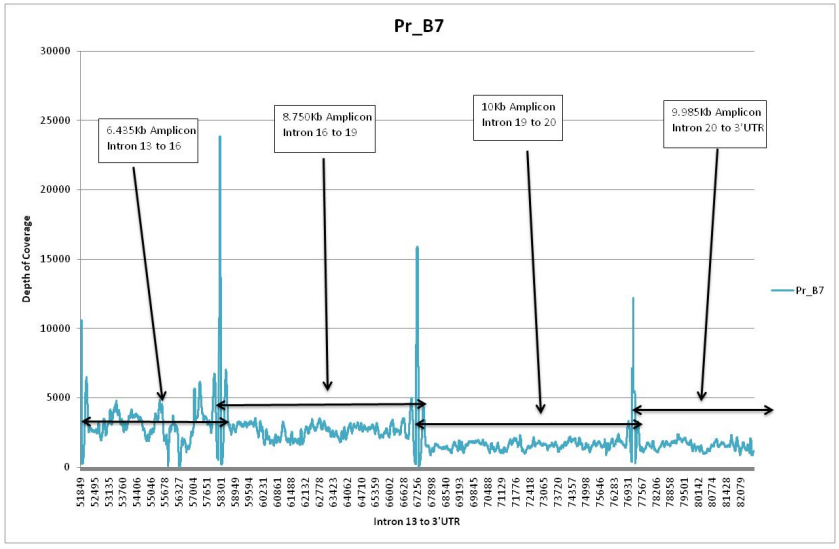

OVOO1.306b
9.5Kb Amplicon
5'UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Depth of Coverage
OVOO1.306b
5'UTR to Intron 6


OVOO1.306b
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 to 13
Depth of Coverage
OVOO1.306b
Intron 6 to Intron 13


OVOO1.306b
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Depth of Coverage
OVOO1.306b
Intron 13 to 3'UTR


Pr_B7
9.5Kb Amplicon
5'UTR to Intron 2
9.58Kb Amplicon
Intron 2 to 3
5.008Kb Amplicon
Intron 3 to 6
Depth of Coverage
5'UTR to Intron 6
Pr_B7


Pr_B7
5.77Kb Amplicon
Intron 6 to 8
8.414Kb Amplicon
Intron 7 to Exon 11
9.411Kb Amplicon
Exon 11 to 13
6.021Kb Amplicon
Intron 12 ro 13
Depth of Coverage
Intron 6 to Intron 13
Pr_B7


Pr_B7
6.435Kb Amplicon
Intron 13 to 16
8.750Kb Amplicon
Intron 16 to 19
10Kb Amplicon
Intron 19 to 20
9.985Kb Amplicon
Intron 20 to 3'UTR
Depth of Coverage
Intron 13 to 3'UTR
Pr_B7

Appendix IV

Genomic co-ordinates of amplicons for 6-gene study

The following tables are the genomic co-ordinates for each amplicon in each gene. The genomic co-ordinates refer to human genome build GRCh37/hg19 February 2009.

RAD51B

Gene Amplicon	Chr	Genomic Co-ordinates	
RAD51L1_t1_1	chr14	68290158	68290345
RAD51L1_t1_2	chr14	68290283	68290477
RAD51L1_t10_1	chr14	68944284	68944475
RAD51L1_t11_1	chr14	68963773	68963971
RAD51L1_t2_1	chr14	68292104	68292274
RAD51L1_t2_2	chr14	68292189	68292377
RAD51L1_t3_1	chr14	68301683	68301876
RAD51L1_t3_2	chr14	68301803	68301977
RAD51L1_t4_1	chr14	68331636	68331821
RAD51L1_t4_2	chr14	68331731	68331930
RAD51L1_t5_1	chr14	68352517	68352697
RAD51L1_t5_2	chr14	68352596	68352785
RAD51L1_t6_1	chr14	68353657	68353851
RAD51L1_t6_3	chr14	68353821	68354017
RAD51L1_t6r_2	chr14	68353747	68353920
RAD51L1_t7_1	chr14	68758545	68758738
RAD51L1_t8_1	chr14	68878090	68878285
RAD51L1_t9_1	chr14	68934826	68935022

RAD51C

Gene Amplicon	Chr	Genomic Co-ordinates	
RAD51C_t1_1	chr17	56769915	56770114
RAD51C_t1_2	chr17	56770032	56770219
RAD51C_t2_1	chr17	56772177	56772374
RAD51C_t2_3	chr17	56772293	56772491
RAD51C_t2_4	chr17	56772391	56772583
RAD51C_t3_1	chr17	56773965	56774164
RAD51C_t3_3	chr17	56774099	56774288
RAD51C_t4_1	chr17	56780503	56780687
RAD51C_t4_2	chr17	56780547	56780744
RAD51C_t5_1	chr17	56787127	56787308
RAD51C_t5_2	chr17	56787219	56787398
RAD51C_t6_1	chr17	56798064	56798180
RAD51C_t6_2	chr17	56798116	56798237
RAD51C_t7_1	chr17	56801334	56801532
RAD51C_t8_1	chr17	56809766	56809950
RAD51C_t9_1	chr17	56811368	56811546
RAD51C_t9_2	chr17	56811471	56811649

RAD51D

Gene Amplicon	Chr	Genomic Co-ordinates	
RAD51L3_t1_1	chr17	33446493	33446664
RAD51L3_t10_1	chr17	33428158	33428357
RAD51L3_t10_2	chr17	33428224	33428419
RAD51L3_t11_1	chr17	33427905	33428098
RAD51L3_t2_1	chr17	33446066	33446256
RAD51L3_t3_1	chr17	33445439	33445618
RAD51L3_t3_2	chr17	33445486	33445684
RAD51L3_t4_10	chr17	33443921	33444094
RAD51L3_T4r_6	chr17	33443828	33444000
RAD51L3_t5_1	chr17	33434328	33434525
RAD51L3_t6_1	chr17	33433949	33434138
RAD51L3_t6_2	chr17	33433991	33434187
RAD51L3_t7_1	chr17	33433343	33433530
RAD51L3_t8_1	chr17	33430401	33430599
RAD51L3_t9_1	chr17	33430193	33430372

XRCC2

Gene Amplicon	Chr	Genomic Co-ordinates	
XRCC2_t1_1	chr7	152373031	152373227
XRCC2_t2_1	chr7	152357715	152357901
XRCC2_t3_1	chr7	152345680	152345871
XRCC2_t3_1r_1	chr7	152345730	152345916
XRCC2_t3_1r_3	chr7	152345919	152346111
XRCC2_t3_4	chr7	152345847	152346021
XRCC2_t3_6	chr7	152346023	152346203
XRCC2_t3_6r_1	chr7	152346086	152346283
XRCC2_t3_8	chr7	152346180	152346364
XRCC2_t3_9	chr7	152346259	152346438
XRCC2_t3s_1	chr7	152346296	152346494

XRCC3

Gene Amplicon	Chr	Genomic Co-ordinates	
XRCC3_t1_1	chr14	104177316	104177500
XRCC3_t2_1	chr14	104174802	104174999
XRCC3_t2_2	chr14	104174852	104175044
XRCC3_t3_1	chr14	104173272	104173461
XRCC3_t3_2	chr14	104173402	104173579
XRCC3_t4_1	chr14	104169454	104169641
XRCC3_t4_2	chr14	104169494	104169690
XRCC3_t5_1	chr14	104165642	104165828
XRCC3_t5_2	chr14	104165758	104165947
XRCC3_t6_1	chr14	104165350	104165547
XRCC3_t7_1	chr14	104165055	104165234
XRCC3_t7_3	chr14	104165214	104165390
XRCC3_t7r_2	chr14	104165179	104165355

SLX4

Gene Amplicon	Chr	Genomic Co-ordinates	
SLX4_13r_3	chr16	3634763	3634958
SLX4_t10_1	chr16	3644414	3644589
SLX4_t10_3	chr16	3644479	3644667
SLX4_t11_1	chr16	3642631	3642817
SLX4_t11_2	chr16	3642733	3642931
SLX4_t12_1	chr16	3638929	3639127
SLX4_t12_10	chr16	3639820	3640019
SLX4_t12_10r_2	chr16	3639919	3640117
SLX4_t12_13	chr16	3639997	3640181
SLX4_t12_14	chr16	3640094	3640276
SLX4_t12_15	chr16	3640185	3640381
SLX4_t12_16	chr16	3640286	3640480
SLX4_t12_17	chr16	3640387	3640568
SLX4_t12_17r_1	chr16	3640475	3640656
SLX4_t12_17r_2	chr16	3640583	3640768
SLX4_t12_2	chr16	3639009	3639203
SLX4_t12_20	chr16	3640687	3640859
SLX4_t12_21	chr16	3640778	3640952
SLX4_t12_22	chr16	3640864	3641063
SLX4_t12_23	chr16	3640965	3641155
SLX4_t12_23r_1	chr16	3641025	3641210
SLX4_t12_23r_2	chr16	3641142	3641341
SLX4_t12_23r_3	chr16	3641226	3641425
SLX4_t12_9	chr16	3639719	3639896
SLX4_t12s_1	chr16	3639108	3639299
SLX4_t12s_2	chr16	3639209	3639394
SLX4_t12s_3	chr16	3639307	3639497
SLX4_t12s_4	chr16	3639401	3639587
SLX4_t12s_5	chr16	3639519	3639718
SLX4_t12s_6	chr16	3639614	3639790
SLX4_t13_1	chr16	3633008	3633195
SLX4_t13_2	chr16	3633116	3633286
SLX4_t13_3	chr16	3633202	3633382
SLX4_t13_4	chr16	3633301	3633480
SLX4_t13_5	chr16	3633395	3633583
SLX4_t13s_1	chr16	3634661	3634859
SLX4_t14_1	chr16	3632272	3632444
SLX4_t14_2	chr16	3632346	3632538
SLX4_t14_3	chr16	3632465	3632664
SLX4_t14_4	chr16	3632556	3632726
SLX4_t2_29	chr16	3658392	3658564
SLX4_t2_33	chr16	3658543	3658742
SLX4_t2_34	chr16	3658660	3658835
SLX4_t2_35	chr16	3658743	3658941
SLX4_t2_37	chr16	3658897	3659079
SLX4_t2r_2	chr16	3658507	3658692
SLX4_t3_1	chr16	3656329	3656526
SLX4_t3_2	chr16	3656444	3656623
SLX4_t3_3	chr16	3656545	3656740
SLX4_t3_4	chr16	3656647	3656846

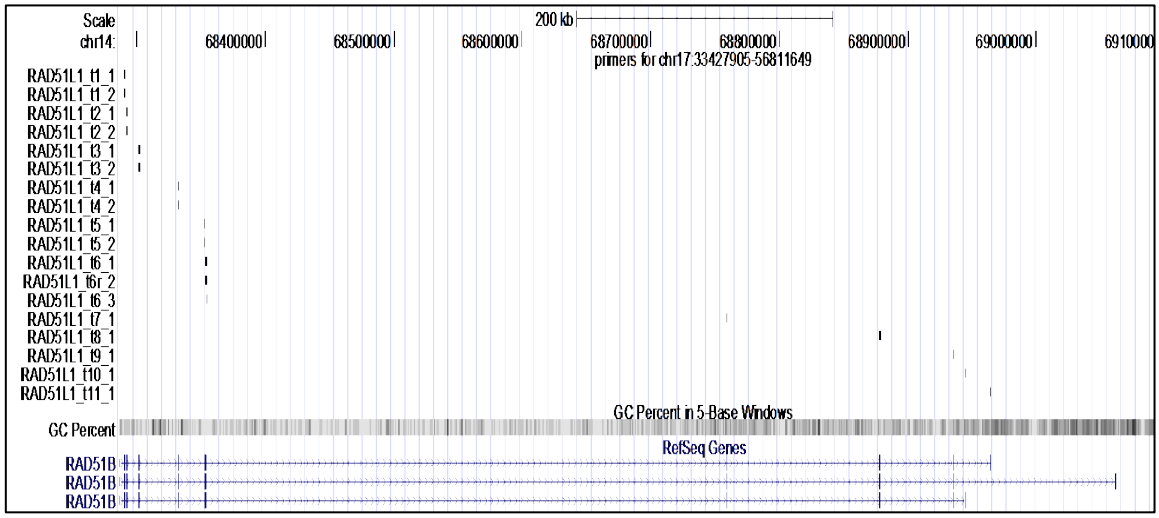
SLX4_t4_1	chr16	3652019	3652216
SLX4_t4_3	chr16	3652190	3652362
SLX4_t4r_2	chr16	3652126	3652319
SLX4_t5_1	chr16	3650883	3651078
SLX4_t5_2	chr16	3650982	3651171
SLX4_t5_3	chr16	3651083	3651257
SLX4_t7_1	chr16	3647264	3647447
SLX4_t7_2	chr16	3647351	3647550
SLX4_t7_3	chr16	3647464	3647646
SLX4_t7_4	chr16	3647556	3647727
SLX4_t7_5	chr16	3647659	3647829
SLX4_t7_6	chr16	3647750	3647935
SLX4_t7_7	chr16	3647858	3648033
SLX4_t7_8	chr16	3647946	3648116
SLX4_t8_1	chr16	3646027	3646226
SLX4_t8_3	chr16	3646204	3646403
SLX4_t8_4	chr16	3646292	3646484
SLX4_t8r_2	chr16	3646147	3646344
SLX4_t9_1	chr16	3645475	3645671
SLX4_t9_2	chr16	3645586	3645760

Appendix V

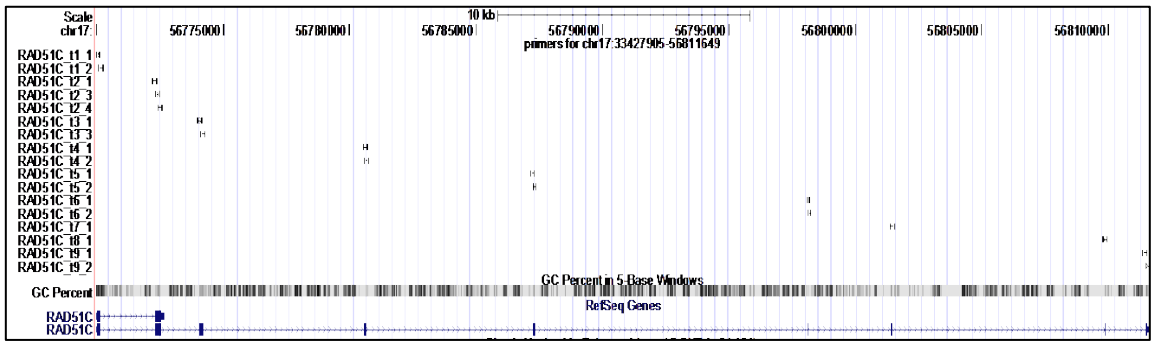
Amplicon maps of target regions for 6-gene study

The following diagrams are amplicon maps of target regions for each gene

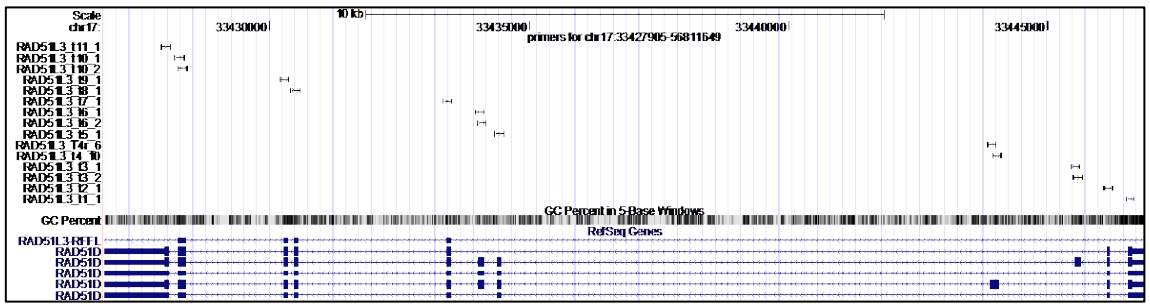
RAD51B



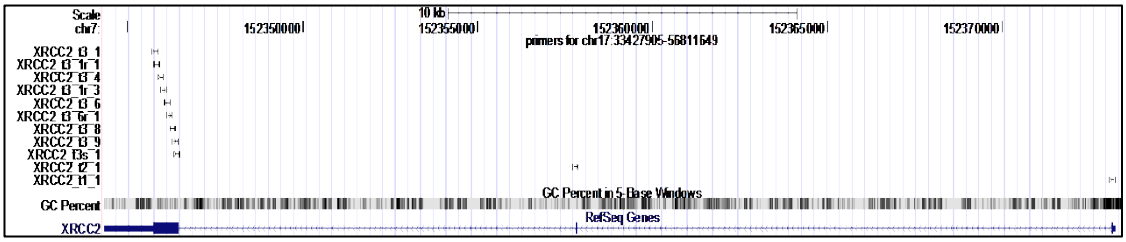
RAD51C



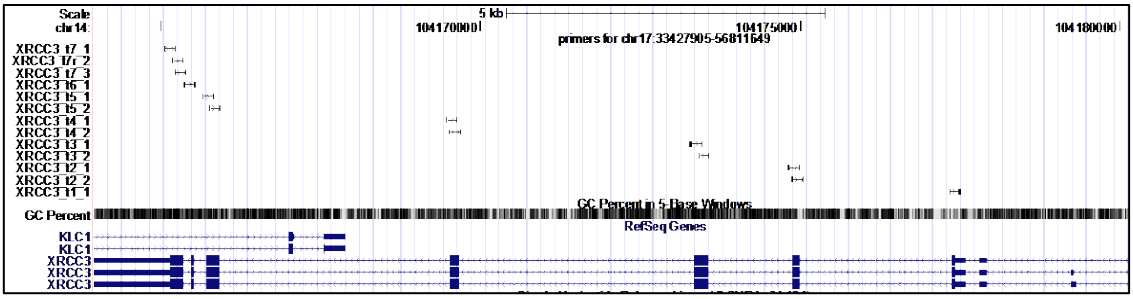
RAD51D



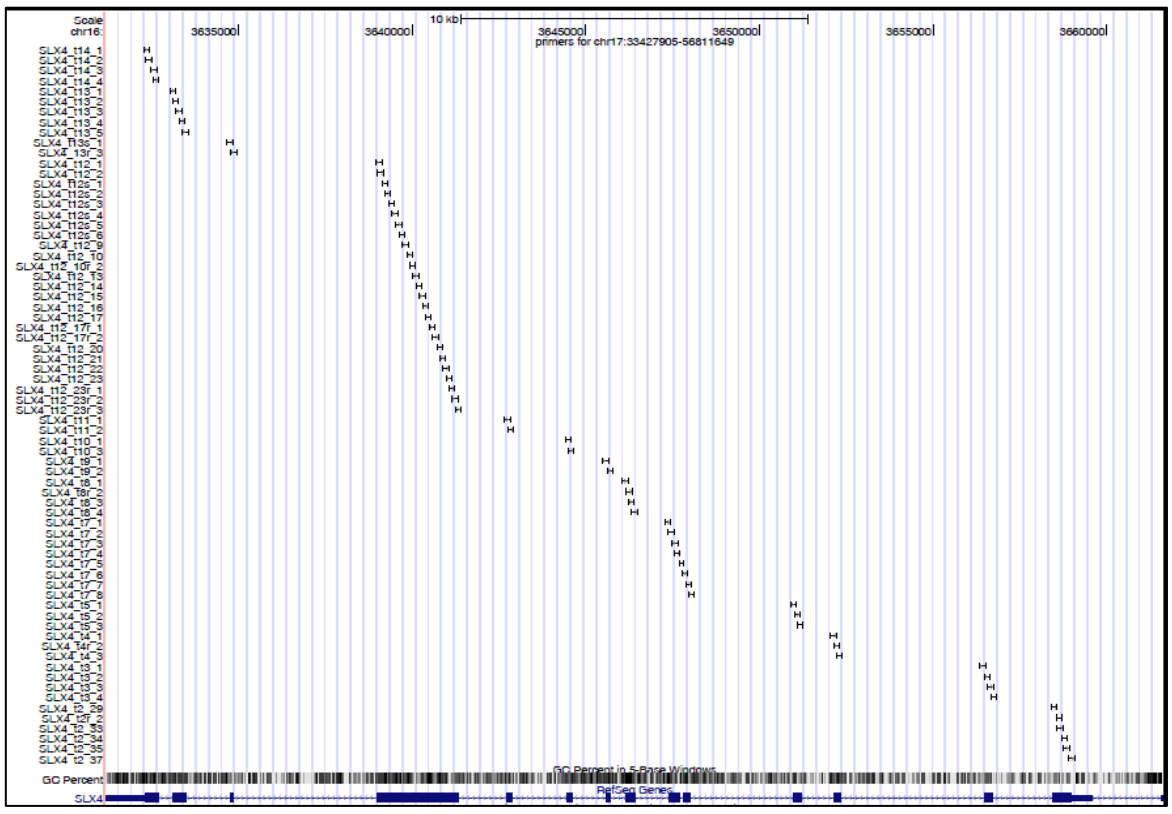
XRCC2



XRCC3



SLX4

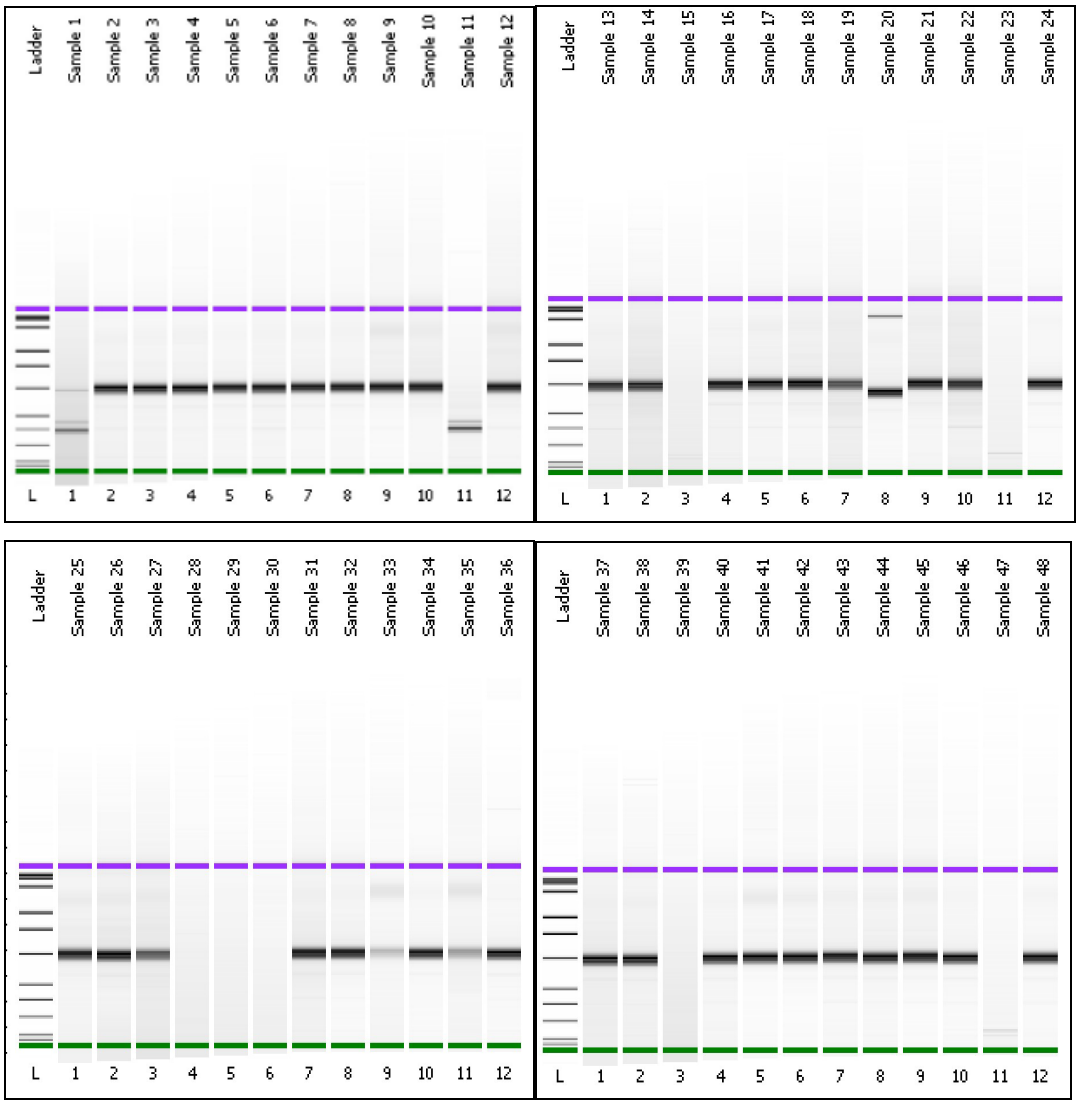


Appendix VI

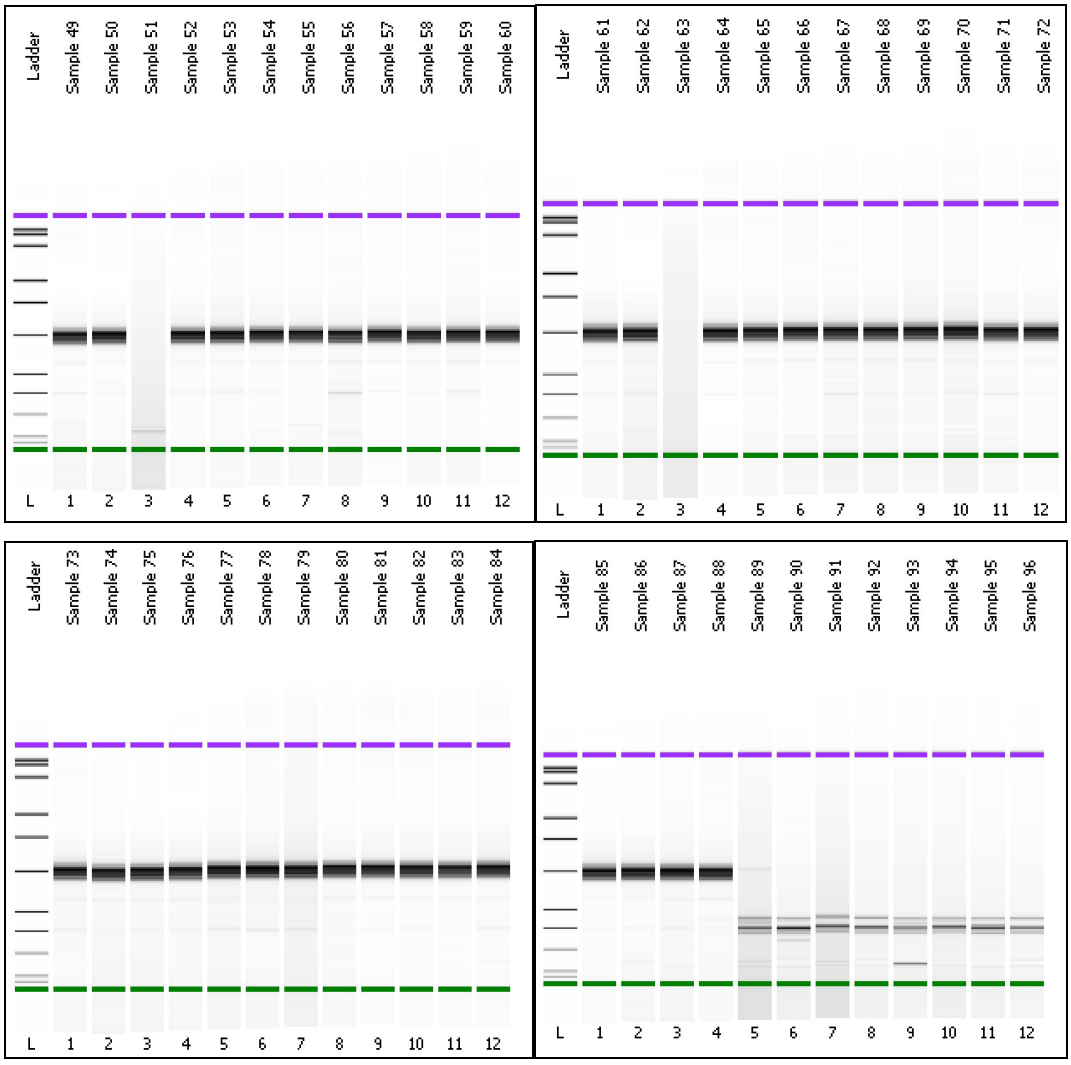
Fluidigm Access Array PCR (6 gene study) results

Agilent Bioanalyzer 2100 results

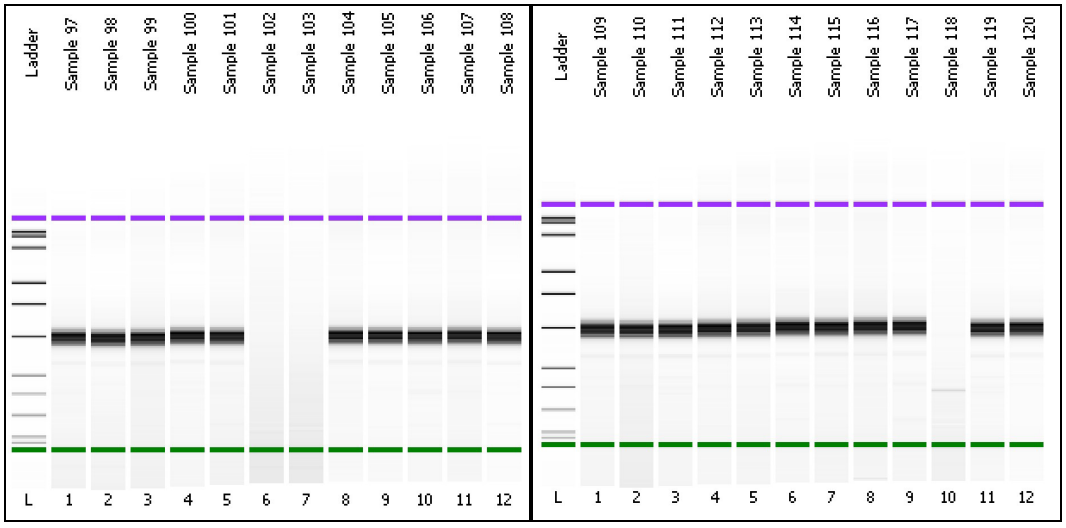
IFC 1

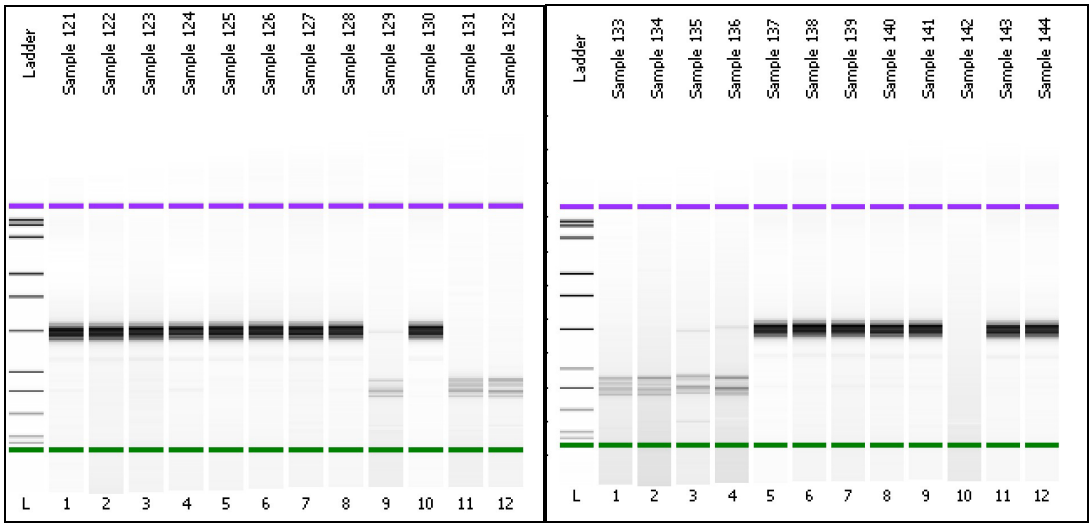


IFC 2

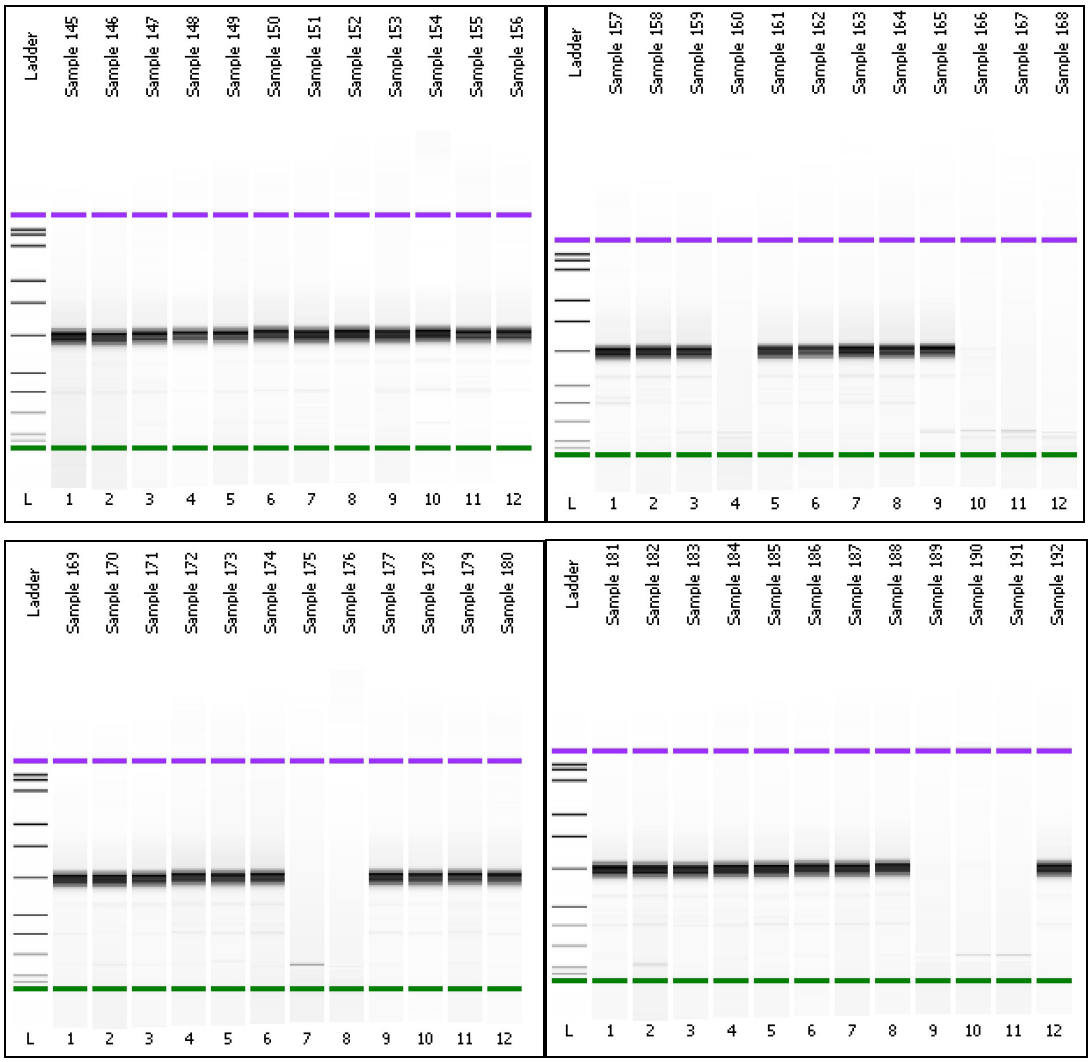

Ladder
Sample 49
Sample 50
Sample 51
Sample 52
Sample 53
Sample 54
Sample 55
Sample 56
Sample 57
Sample 58
Sample 59
Sample 60
Sample 61
Sample 62
Sample 63
Sample 64
Sample 65
Sample 66
Sample 67
Sample 68
Sample 69
Sample 70
Sample 71
Sample 72
Sample 73
Sample 74
Sample 75
Sample 76
Sample 77
Sample 78
Sample 79
Sample 80
Sample 81
Sample 82
Sample 83
Sample 84
Sample 85
Sample 86
Sample 87
Sample 88
Sample 89
Sample 90
Sample 91
Sample 92
Sample 93
Sample 94
Sample 95
Sample 96
L 1 2 3 4 5 6 7 8 9 10 11 12

IFC 3

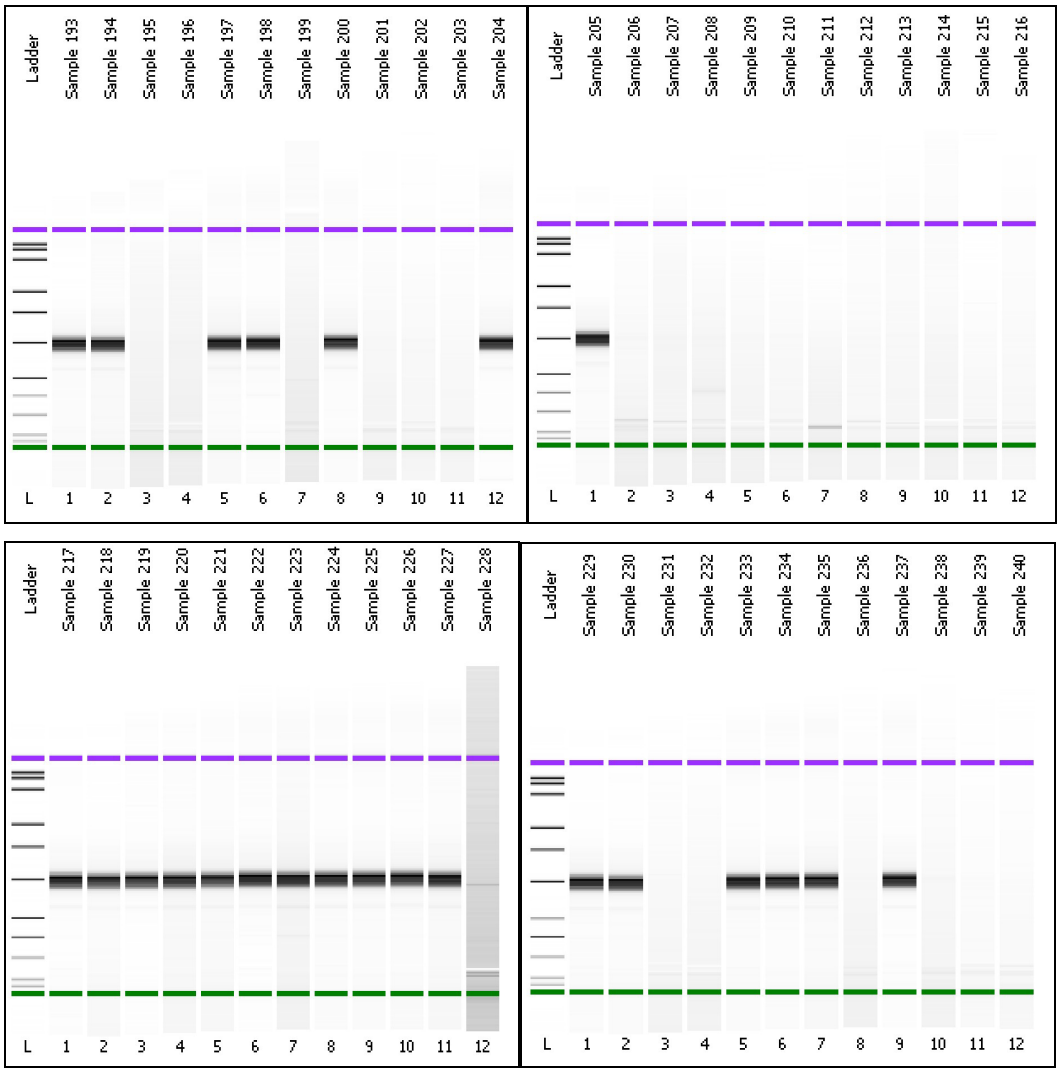

Ladder
Sample 97
Sample 98
Sample 99
Sample 100
Sample 101
Sample 102
Sample 103
Sample 104
Sample 105
Sample 106
Sample 107
Sample 108
Sample 109
Sample 110
Sample 111
Sample 112
Sample 113
Sample 114
Sample 115
Sample 116
Sample 117
Sample 118
Sample 119
Sample 120
L 1 2 3 4 5 6 7 8 9 10 11 12


Ladder
Sample 121
Sample 122
Sample 123
Sample 124
Sample 125
Sample 126
Sample 127
Sample 128
Sample 129
Sample 130
Sample 131
Sample 132
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 133
Sample 134
Sample 135
Sample 136
Sample 137
Sample 138
Sample 139
Sample 140
Sample 141
Sample 142
Sample 143
Sample 144
L 1 2 3 4 5 6 7 8 9 10 11 12

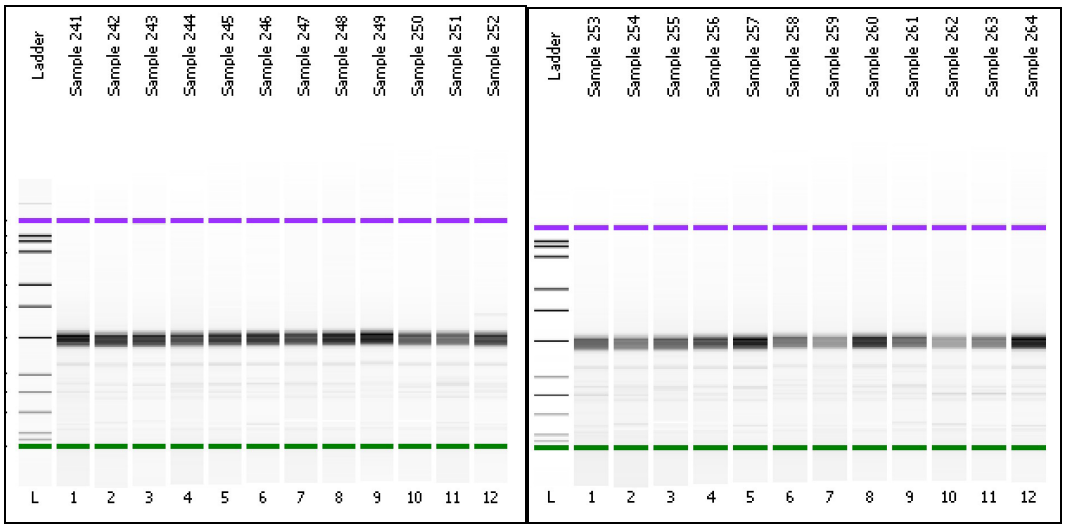
IFC 4

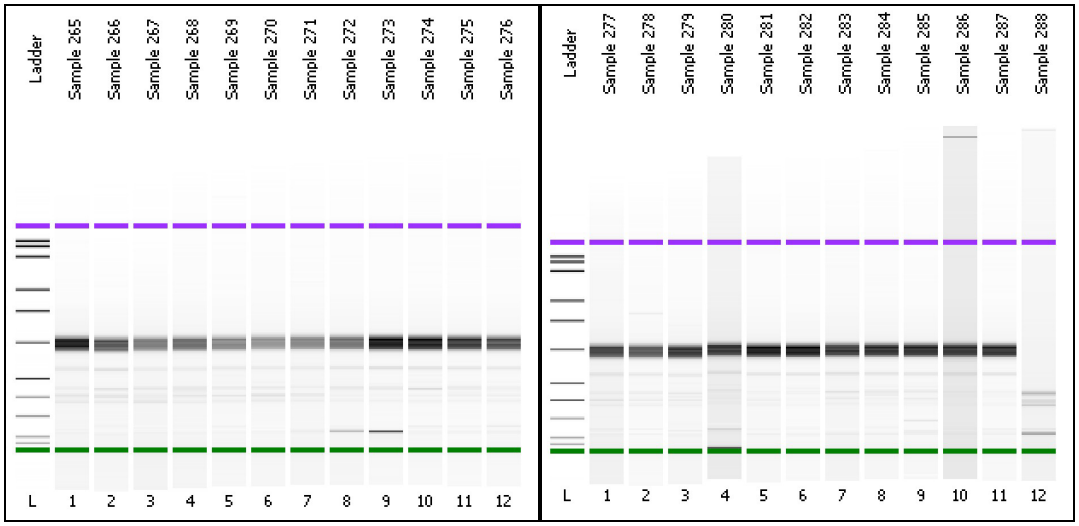

Ladder
Sample 145
Sample 146
Sample 147
Sample 148
Sample 149
Sample 150
Sample 151
Sample 152
Sample 153
Sample 154
Sample 155
Sample 156
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 157
Sample 158
Sample 159
Sample 160
Sample 161
Sample 162
Sample 163
Sample 164
Sample 165
Sample 166
Sample 167
Sample 168
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 169
Sample 170
Sample 171
Sample 172
Sample 173
Sample 174
Sample 175
Sample 176
Sample 177
Sample 178
Sample 179
Sample 180
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 181
Sample 182
Sample 183
Sample 184
Sample 185
Sample 186
Sample 187
Sample 188
Sample 189
Sample 190
Sample 191
Sample 192
L 1 2 3 4 5 6 7 8 9 10 11 12

IFC 5

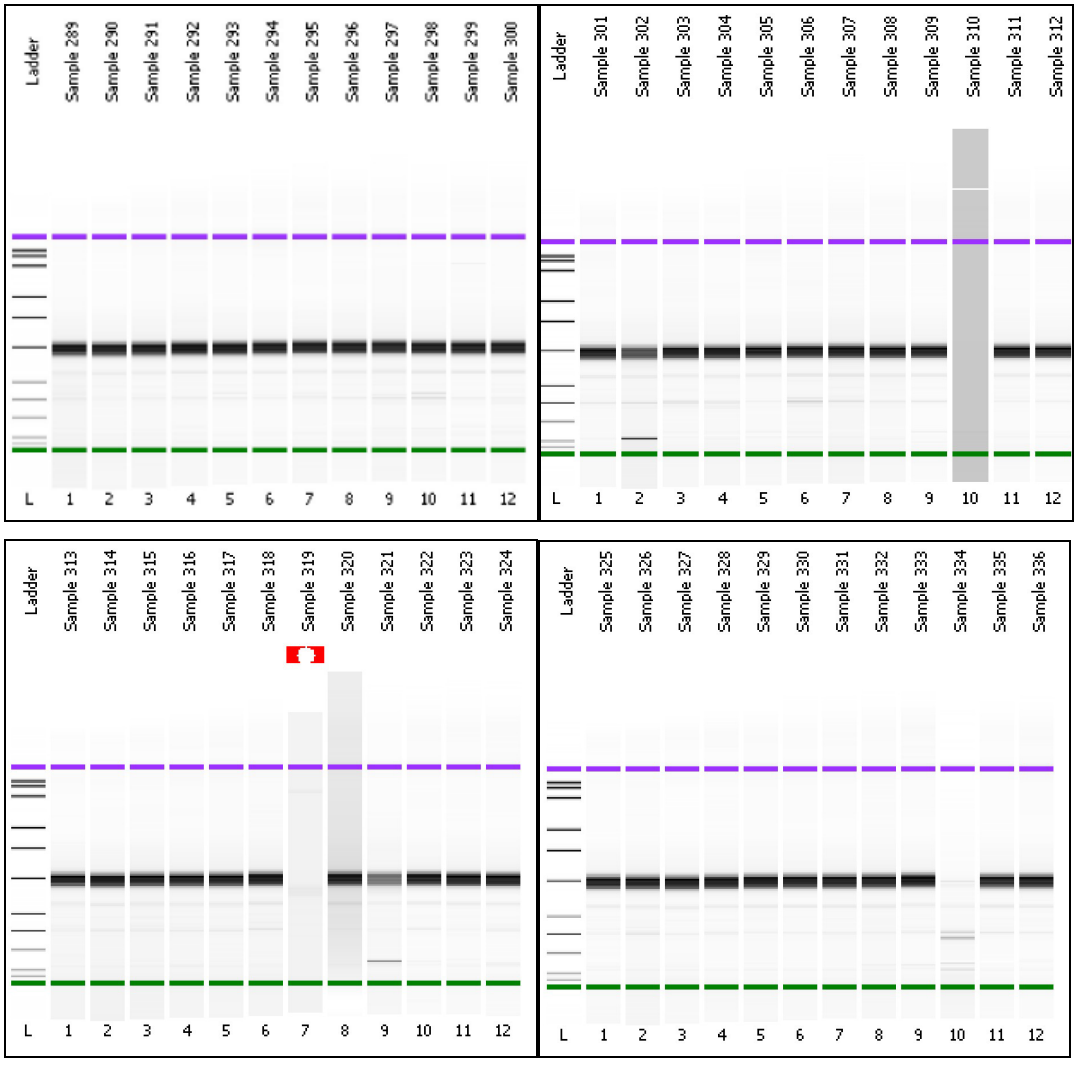

Ladder
Sample 193
Sample 194
Sample 195
Sample 196
Sample 197
Sample 198
Sample 199
Sample 200
Sample 201
Sample 202
Sample 203
Sample 204
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 205
Sample 206
Sample 207
Sample 208
Sample 209
Sample 210
Sample 211
Sample 212
Sample 213
Sample 214
Sample 215
Sample 216
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 217
Sample 218
Sample 219
Sample 220
Sample 221
Sample 222
Sample 223
Sample 224
Sample 225
Sample 226
Sample 227
Sample 228
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 229
Sample 230
Sample 231
Sample 232
Sample 233
Sample 234
Sample 235
Sample 236
Sample 237
Sample 238
Sample 239
Sample 240
L 1 2 3 4 5 6 7 8 9 10 11 12

IFC 6

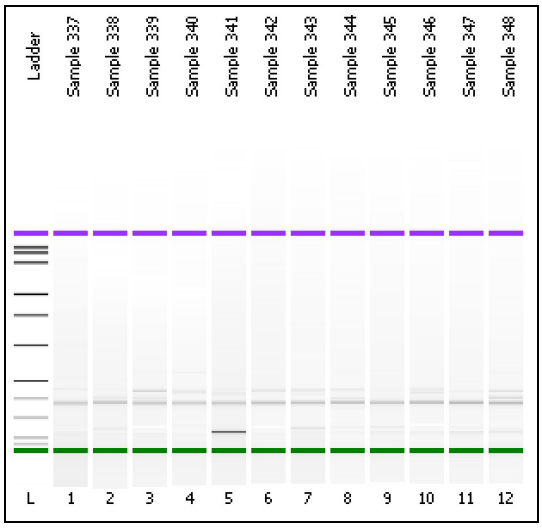

Ladder
Sample 241
Sample 242
Sample 243
Sample 244
Sample 245
Sample 246
Sample 247
Sample 248
Sample 249
Sample 250
Sample 251
Sample 252
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 253
Sample 254
Sample 255
Sample 256
Sample 257
Sample 258
Sample 259
Sample 260
Sample 261
Sample 262
Sample 263
Sample 264
L 1 2 3 4 5 6 7 8 9 10 11 12


Ladder
Sample 265
Sample 266
Sample 267
Sample 268
Sample 269
Sample 270
Sample 271
Sample 272
Sample 273
Sample 274
Sample 275
Sample 276
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 277
Sample 278
Sample 279
Sample 280
Sample 281
Sample 282
Sample 283
Sample 284
Sample 285
Sample 286
Sample 287
Sample 288
L 1 2 3 4 5 6 7 8 9 10 11 12

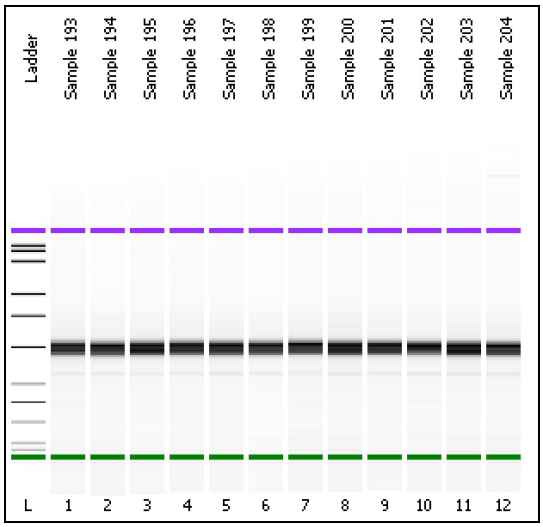
IFC 7

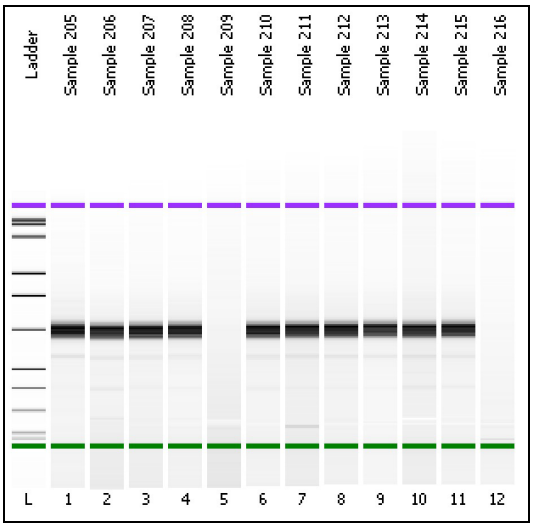

Ladder
Sample 289
Sample 290
Sample 291
Sample 292
Sample 293
Sample 294
Sample 295
Sample 296
Sample 297
Sample 298
Sample 299
Sample 300
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 301
Sample 302
Sample 303
Sample 304
Sample 305
Sample 306
Sample 307
Sample 308
Sample 309
Sample 310
Sample 311
Sample 312
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 313
Sample 314
Sample 315
Sample 316
Sample 317
Sample 318
Sample 319
Sample 320
Sample 321
Sample 322
Sample 323
Sample 324
L 1 2 3 4 5 6 7 8 9 10 11 12
Ladder
Sample 325
Sample 326
Sample 327
Sample 328
Sample 329
Sample 330
Sample 331
Sample 332
Sample 333
Sample 334
Sample 335
Sample 336
L 1 2 3 4 5 6 7 8 9 10 11 12

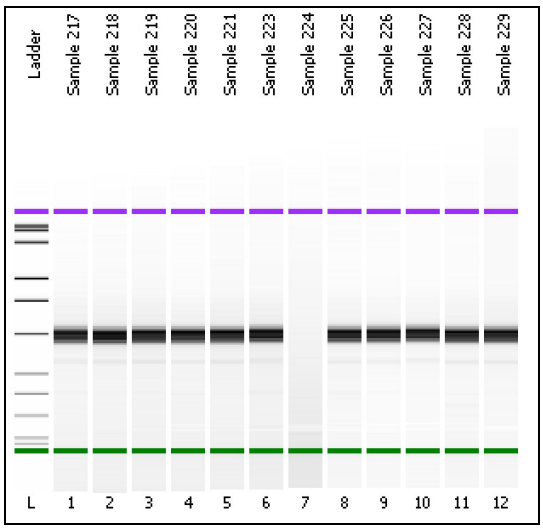
IFC 8

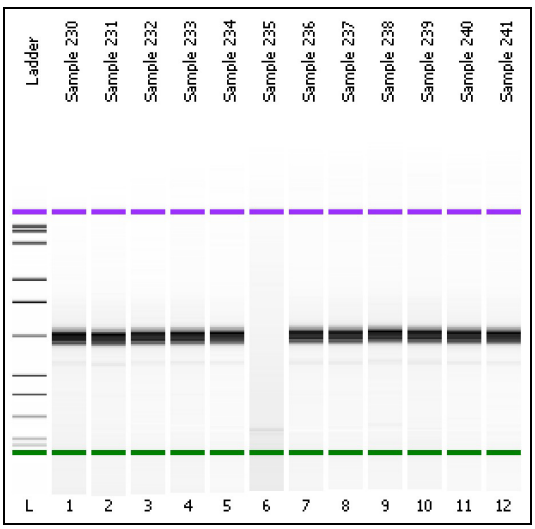


IFC 5 (Repeat)

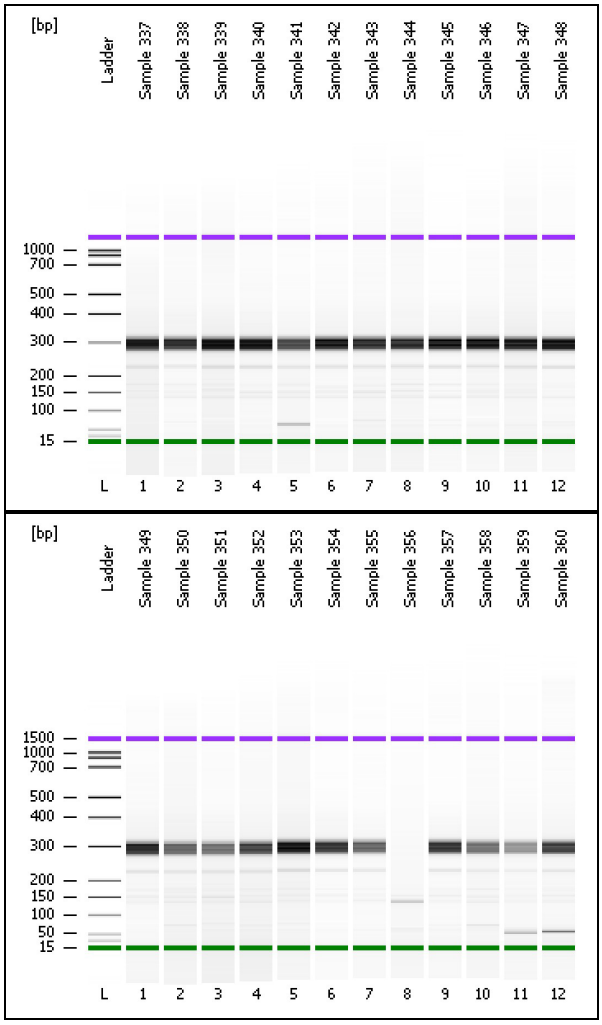


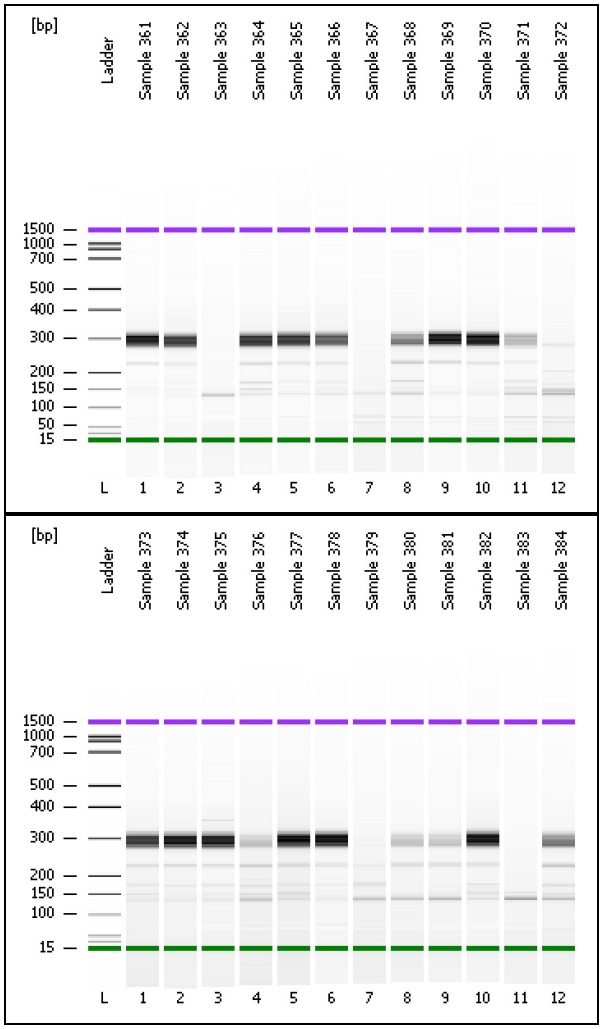






IFC 9


[bp]
Ladder
Sample 337
Sample 338
Sample 339
Sample 340
Sample 341
Sample 342
Sample 343
Sample 344
Sample 345
Sample 346
Sample 347
Sample 348
1000
700
500
400
300
200
150
100
15
L 1 2 3 4 5 6 7 8 9 10 11 12
[bp]
Ladder
Sample 349
Sample 350
Sample 351
Sample 352
Sample 353
Sample 354
Sample 355
Sample 356
Sample 357
Sample 358
Sample 359
Sample 360
1500
1000
700
500
400
300
200
150
100
50
15
L 1 2 3 4 5 6 7 8 9 10 11 12


[bp]
Ladder
Sample 361
Sample 362
Sample 363
Sample 364
Sample 365
Sample 366
Sample 367
Sample 368
Sample 369
Sample 370
Sample 371
Sample 372
1500
1000
700
500
400
300
200
150
100
50
15
L
1
2
3
4
5
6
7
8
9
10
11
12
[bp]
Ladder
Sample 373
Sample 374
Sample 375
Sample 376
Sample 377
Sample 378
Sample 379
Sample 380
Sample 381
Sample 382
Sample 383
Sample 384
1500
1000
700
500
400
300
200
150
100
15
L
1
2
3
4
5
6
7
8
9
10
11
12

Appendix VII

Final concentration and dilutions for each lane

ID	Initial Molarity	Vol DNA	Vol H_2O	Required Final Molarity
Lane 1 Pool	211.1	2.37	47.63	10 nM
Lane 2 Pool	74.7	6.7	43.3	10 nM
Lane 3 Pool	71.8	6.9	43.1	10 nM
Lane 4 Pool	196.3	2.58	47.42	10 nM
Lane 5 Pool	44.56	11.23	38.77	10 nm
Lane 6 Pool	201.4	2.48	47.52	10 nM
Lane 7 Pool	81.1	6.17	43.83	10 nM

Appendix VIII

Dilution and final QC for flow cell

Once diluted each final pool is quantified in triplicate using the Agilent Bioanalyzer. The mean is calculated and used as the actual final molarity for pools

Lane ID	Final Molarity (nM)
Lane 1	7.0
Lane 2	11.2
Lane 3	9.5
Lane 4	8.1
Lane 5	8.7
Lane 6	11.4
Lane 7	9.9

Appendix IX

Genomic co-ordinates of amplicons for additional genes for the 9-gene study

The following tables and images give the genomic co-ordinates for each amplicon. Genomic co-ordinates refer to human genome build GRCh37/hg19 February 2009

BRCA1

Amplicon	chr	Start	End
BRCA1_ex2_1	chr17	41275935	41276143
BRCA1_ex3_1	chr17	41267715	41267905
BRCA1_ex5_1	chr17	41258424	41258631
BRCA1_ex6_1	chr17	41256828	41257012
BRCA1_ex7_1	chr17	41256179	41256380
BRCA1_ex7_2	chr17	41256125	41256330
BRCA1_ex8_1	chr17	41251822	41252016
BRCA1_ex8_2	chr17	41251728	41251916
BRCA1_ex9_1	chr17	41249269	41249472
BRCA1_ex9_1r	chr17	41249238	41249431
BRCA1_ex9_2	chr17	41249204	41249333
BRCA1_ex10_1	chr17	41247812	41248010
BRCA1_ex11_1	chr17	41246725	41246923
BRCA1_ex11_2	chr17	41246611	41246816
BRCA1_ex11_3	chr17	41246489	41246698
BRCA1_ex11_4	chr17	41246432	41246591
BRCA1_ex11_5	chr17	41246346	41246517
BRCA1_ex11_6	chr17	41246248	41246455
BRCA1_ex11_7	chr17	41246143	41246351
BRCA1_ex11_8	chr17	41246056	41246217
BRCA1_ex11_9	chr17	41245978	41246135
BRCA1_ex11_10	chr17	41245825	41246030
BRCA1_ex11_11	chr17	41245726	41245932
BRCA1_ex11_12	chr17	41245629	41245813
BRCA1_ex11_13	chr17	41245504	41245713
BRCA1_ex11_14	chr17	41245430	41245630
BRCA1_ex11_15	chr17	41245358	41245542
BRCA1_ex11_16	chr17	41245207	41245410
BRCA1_ex11_17	chr17	41245129	41245318
BRCA1_ex11_18	chr17	41245011	41245216
BRCA1_ex11_19	chr17	41244921	41245120
BRCA1_ex11_20	chr17	41244805	41245000
BRCA1_ex11_21	chr17	41244689	41244891
BRCA1_ex11_22	chr17	41244600	41244786
BRCA1_ex11_23	chr17	41244473	41244682
BRCA1_ex11_24	chr17	41244417	41244614
BRCA1_ex11_25	chr17	41244293	41244481
BRCA1_ex11_26	chr17	41244213	41244402
BRCA1_ex11_27	chr17	41244102	41244295
BRCA1_ex11_28	chr17	41243995	41244197

BRCA1_ex11_29	chr17	41243875	41244080
BRCA1_ex11_31	chr17	41243852	41243983
BRCA1_ex11_30	chr17	41243758	41243898
BRCA1_ex11_32	chr17	41243670	41243879
BRCA1_ex11_33	chr17	41243604	41243793
BRCA1_ex11_34	chr17	41243455	41243655
BRCA1_ex11_35	chr17	41243368	41243567
BRCA1_ex12_1	chr17	41242904	41243110
BRCA1_ex13_1	chr17	41234497	41234646
BRCA1_ex13_2r	chr17	41234400	41234556
BRCA1_ex13_2	chr17	41234369	41234566
BRCA1_in13_1	chr17	41231312	41231476
BRCA1_ex14_1	chr17	41228464	41228653
BRCA1_ex15_1	chr17	41226443	41226609
BRCA1_ex15_2r	chr17	41226316	41226503
BRCA1_ex16_1r	chr17	41223199	41223398
BRCA1_ex16_1	chr17	41223130	41223311
BRCA1_ex16_2	chr17	41223050	41223251
BRCA1_ex16_3	chr17	41222901	41223096
BRCA1_ex17_1	chr17	41219567	41219776
BRCA1_ex18_1	chr17	41215840	41216029
BRCA1_ex19_1	chr17	41215258	41215442
BRCA1_ex20_1	chr17	41209045	41209232
BRCA1_ex21_1	chr17	41203032	41203166
BRCA1_ex22_1	chr17	41201071	41201275
BRCA1_ex23_1	chr17	41199621	41199755
BRCA1_ex24_1	chr17	41197714	41197895
BRCA1_ex24_2	chr17	41197574	41197757

BRCA2

Amplicon	chr	Start	End
BRCA2_ex1_1	chr13	32890475	32890673
BRCA2_ex1_2	chr13	32890571	32890753
BRCA2_ex2_1	chr13	32893173	32893381
BRCA2_ex2_2	chr13	32893249	32893426
BRCA2_ex3_1	chr13	32899205	32899334
BRCA2_ex4_1	chr13	32900202	32900400
BRCA2_ex6_1	chr13	32900582	32900786
BRCA2_ex7_1	chr13	32903475	32903668
BRCA2_ex8_1	chr13	32904995	32905180
BRCA2_ex9_1	chr13	32906365	32906514
BRCA2_ex9_2	chr13	32906405	32906604
BRCA2_ex9_3	chr13	32906512	32906710
BRCA2_ex9_4	chr13	32906612	32906786
BRCA2_ex9_5	chr13	32906717	32906879
BRCA2_ex9_6	chr13	32906760	32906969
BRCA2_ex9_7	chr13	32906848	32907051

BRCA2_ex9_8	chr13	32906967	32907169
BRCA2_ex9_9	chr13	32907062	32907270
BRCA2_ex9_10	chr13	32907219	32907384
BRCA2_ex9_11	chr13	32907326	32907512
BRCA2_ex9_11r	chr13	32907328	32907512
BRCA2_ex9_12	chr13	32907462	32907592
BRCA2_ex10_1	chr13	32910341	32910505
BRCA2_ex10_2	chr13	32910454	32910654
BRCA2_ex10_3	chr13	32910583	32910749
BRCA2_ex10_4	chr13	32910671	32910857
BRCA2_ex10_5	chr13	32910787	32910992
BRCA2_ex10_5r	chr13	32910803	32911001
BRCA2_ex10_6	chr13	32910906	32911074
BRCA2_ex10_7	chr13	32910983	32911134
BRCA2_ex10_8	chr13	32911192	32911396
BRCA2_ex10_9	chr13	32911249	32911431
BRCA2_ex10_10	chr13	32911360	32911557
BRCA2_ex10_10r	chr13	32911339	32911533
BRCA2_ex10_11	chr13	32911509	32911678
BRCA2_ex10_12	chr13	32911620	32911821
BRCA2_ex10_12r	chr13	32911622	32911812
BRCA2_ex10_13	chr13	32911753	32911925
BRCA2_ex10_14	chr13	32911796	32911940
BRCA2_ex10_15	chr13	32911888	32912028
BRCA2_ex10_16	chr13	32911986	32912151
BRCA2_ex10_17	chr13	32912045	32912250
BRCA2_ex10_18	chr13	32912247	32912443
BRCA2_ex10_19	chr13	32912285	32912479
BRCA2_ex10_20	chr13	32912423	32912571
BRCA2_ex10_20r1	chr13	32912467	32912630
BRCA2_ex10_20r2	chr13	32912384	32912516
BRCA2_ex10_21	chr13	32912499	32912692
BRCA2_ex10_22	chr13	32912642	32912809
BRCA2_ex10_23	chr13	32912716	32912877
BRCA2_ex10_24	chr13	32912830	32913015
BRCA2_ex10_25	chr13	32912916	32913078
BRCA2_ex10_26	chr13	32912977	32913156
BRCA2_ex10_27	chr13	32913018	32913197
BRCA2_ex10_28	chr13	32913133	32913333
BRCA2_ex10_29	chr13	32913250	32913382
BRCA2_ex10_30	chr13	32913318	32913454
BRCA2_ex10_31	chr13	32913396	32913598
BRCA2_ex10_32	chr13	32913491	32913669
BRCA2_ex10_33	chr13	32913583	32913759
BRCA2_ex10_33r	chr13	32913566	32913758

BRCA2_ex10_34	chr13	32913674	32913883
BRCA2_ex10_35	chr13	32913735	32913910
BRCA2_ex10_36	chr13	32913854	32914019
BRCA2_ex10_37	chr13	32913939	32914127
BRCA2_ex10_38	chr13	32914006	32914206
BRCA2_ex10_39	chr13	32914146	32914287
BRCA2_ex10_40	chr13	32914225	32914408
BRCA2_ex10_41	chr13	32914290	32914449
BRCA2_ex10_42	chr13	32914394	32914573
BRCA2_ex10_43	chr13	32914452	32914605
BRCA2_ex10_44	chr13	32914560	32914689
BRCA2_ex10_45	chr13	32914638	32914844
BRCA2_ex10_46	chr13	32914795	32914971
BRCA2_ex10_46r1	chr13	32914827	32914988
BRCA2_ex10_46r2	chr13	32914755	32914924
BRCA2_ex10_47	chr13	32914896	32915061
BRCA2_ex10_48	chr13	32914946	32915095
BRCA2_ex10_49	chr13	32915032	32915240
BRCA2_ex10_50	chr13	32915157	32915364
BRCA2_ex10_51	chr13	32915253	32915461
BRCA2_ex11_1	chr13	32918555	32918763
BRCA2_ex11_2	chr13	32918687	32918851
BRCA2_ex12_1	chr13	32920895	32921044
BRCA2_ex13_1	chr13	32928918	32929127
BRCA2_ex13_2	chr13	32929074	32929253
BRCA2_ex13_3	chr13	32929104	32929295
BRCA2_ex13_4	chr13	32929240	32929397
BRCA2_ex13_5	chr13	32929340	32929536
BRCA2_ex14_1	chr13	32930503	32930636
BRCA2_ex14_2	chr13	32930573	32930742
BRCA2_ex14_3	chr13	32930661	32930870
BRCA2_ex15_1	chr13	32931809	32932007
BRCA2_ex15_2	chr13	32931904	32932113
BRCA2_ex16_1	chr13	32936572	32936763
BRCA2_ex16_2	chr13	32936649	32936821
BRCA2_ex16_3	chr13	32936778	32936939
BRCA2_ex17_1	chr13	32937215	32937401
BRCA2_ex17_1r	chr13	32937263	32937435
BRCA2_ex17_2	chr13	32937347	32937512
BRCA2_ex17_3	chr13	32937452	32937581
BRCA2_ex17_4	chr13	32937537	32937668
BRCA2_ex17_5	chr13	32937594	32937785
BRCA2_ex18_1	chr13	32944422	32944613
BRCA2_ex18_1r	chr13	32944440	32944628
BRCA2_ex18_2	chr13	32944546	32944696

BRCA2_ex18_3	chr13	32944603	32944733
BRCA2_ex19_1	chr13	32945056	32945239
BRCA2_ex19_2	chr13	32945140	32945279
BRCA2_ex20_1	chr13	32950752	32950959
BRCA2_ex21_1	chr13	32953401	32953606
BRCA2_ex21_2	chr13	32953491	32953694
BRCA2_ex22_1	chr13	32953773	32953978
BRCA2_ex22_2	chr13	32953901	32954099
BRCA2_ex22_1r1	chr13	32953850	32954025
BRCA2_ex22_1r2	chr13	32953893	32954060
BRCA2_ex22_1r3	chr13	32954003	32954185
BRCA2_ex23_1	chr13	32954047	32954225
BRCA2_ex23_2	chr13	32954179	32954375
BRCA2_ex24_1	chr13	32968741	32968935
BRCA2_ex24_2	chr13	32968827	32969028
BRCA2_ex24_3	chr13	32968944	32969139
BRCA2_ex25_1	chr13	32970909	32971109
BRCA2_ex25_2	chr13	32971008	32971211
BRCA2_ex26_1	chr13	32972189	32972375
BRCA2_ex26_2	chr13	32972306	32972470
BRCA2_ex26_3	chr13	32972421	32972628
BRCA2_ex26_4	chr13	32972559	32972675
BRCA2_ex26_5	chr13	32972600	32972774
BRCA2_ex26_6	chr13	32972725	32972883
BRCA2_ex26_7	chr13	32972817	32972988
BRCA2_ex5_1	chr13	32900331	32900458

BRIP1

Amplicon	chr	Start	End
BRIP1_EX2_t1	chr17	59938820	59938979
BRIP1_EX2_t2	chr17	59938723	59938909
BRIP1_EX3_t1	chr17	59937113	59937296
BRIP1_EX4_t1	chr17	59934461	59934660
BRIP1_EX4_t2	chr17	59934388	59934533
BRIP1_EX5_t1	chr17	59926496	59926667
BRIP1_EX5_t2	chr17	59926404	59926580
BRIP1_EX6_t1	chr17	59924424	59924617
BRIP1_EX7_t1	chr17	59885975	59886147
BRIP1_EX7_t2	chr17	59885877	59886071
BRIP1_EX7_t3	chr17	59885825	59885997
BRIP1_EX7_t4	chr17	59885725	59885923
BRIP1_EX8_t1	chr17	59878683	59878879
BRIP1_EX8_t2	chr17	59878566	59878762
BRIP1_EX9_t1	chr17	59876535	59876721

BRIP1_EX9_t2	chr17	59876415	59876605
BRIP1_EX10_t1	chr17	59870923	59871122
BRIP1_EX11_t1	chr17	59861686	59861836
BRIP1_EX11_t2	chr17	59861584	59861774
BRIP1_EX12_t1	chr17	59858238	59858416
BRIP1_EX12_t2	chr17	59858152	59858336
BRIP1_EX13_t1	chr17	59857651	59857817
BRIP1_EX13_t2	chr17	59857611	59857776
BRIP1_EX14_t1	chr17	59853789	59853963
BRIP1_EX14_t2	chr17	59853694	59853886
BRIP1_EX15_t1	chr17	59821856	59822044
BRIP1_EX15_t2	chr17	59821760	59821936
BRIP1_EX16_t1	chr17	59820350	59820533
BRIP1_EX17_t1	chr17	59793353	59793548
BRIP1_EX17_t2	chr17	59793292	59793425
BRIP1_EX18_t1	chr17	59770774	59770942
BRIP1_EX18_t2	chr17	59770713	59770860
BRIP1_EX19_t1	chr17	59763388	59763560
BRIP1_EX19_t2	chr17	59763309	59763498
BRIP1_EX19_t3	chr17	59763197	59763395
BRIP1_EX19_t4	chr17	59763152	59763302
BRIP1_EX20_t1	chr17	59761365	59761557
BRIP1_EX20_t2	chr17	59761251	59761449
BRIP1_EX20_t3	chr17	59761155	59761343
BRIP1_EX20_t4	chr17	59761056	59761225
BRIP1_EX20_t5	chr17	59760941	59761140
BRIP1_EX20_t6	chr17	59760847	59761046
BRIP1_EX20_t7	chr17	59760741	59760937
BRIP1_EX20_t8	chr17	59760654	59760824
BRIP1_EX20_t9	chr17	59760555	59760750

PALB2

Amplicon	chr	Start	End
PALB2_EX1_t1	chr16	23652356	23652513
PALB2_EX2_t1	chr16	23649327	23649505
PALB2_EX3_t1	chr16	23649139	23649301
PALB2_EX4_t1	chr16	23647518	23647668
PALB2_EX4_t2	chr16	23647424	23647609
PALB2_EX4_t3	chr16	23647314	23647512
PALB2_EX4_t4	chr16	23647214	23647389
PALB2_EX4_t5	chr16	23647122	23647308
PALB2_EX4_t6	chr16	23646995	23647194
PALB2_EX4_t7	chr16	23646882	23647073
PALB2_EX4_t8	chr16	23646775	23646969

PALB2_EX4_t9	chr16	23646670	23646869
PALB2_EX4_t10	chr16	23646567	23646758
PALB2_EX4_t11	chr16	23646460	23646640
PALB2_EX4_t12	chr16	23646395	23646593
PALB2_EX4_t13	chr16	23646244	23646443
PALB2_EX4_t14	chr16	23646130	23646326
PALB2_EX5_t1	chr16	23641636	23641834
PALB2_EX5_t2	chr16	23641517	23641715
PALB2-EX5_t3	chr16	23641394	23641587
PALB2_EX5_t4	chr16	23641289	23641488
PALB2_EX5_t5	chr16	23641188	23641387
PALB2_EX5_t6	chr16	23641073	23641261
PALB2_EX5_t7	chr16	23640967	23641161
PALB2_EX5_t8	chr16	23640861	23641052
PALB2_EX6_t1	chr16	23640491	23640672
PALB2_EX7_t1	chr16	23637560	23637758
PALB2_EX7_t2	chr16	23637455	23637645
PALB2_EX8_t1	chr16	23635297	23635483
PALB2_EX9_t1	chr16	23634292	23634489
PALB2_EX9_t2	chr16	23634179	23634378
PALB2_EX10_t1	chr16	23632650	23632848
PALB2_EX11_t1	chr16	23625297	23625476
PALB2_EX12_t1	chr16	23619192	23619361
PALB2_EX12_t2	chr16	23619085	23619278
PALB2_EX13_t1	chr16	23614894	23615081
PALB2_EX13_t2	chr16	23614757	23614956
PALB2_EX13_t3	chr16	23614654	23614832

NBN

Amplicon	Chr	Start	End
NBN_EX1_t1	chr8	90996705	90996847
NBN_EX2_t1	chr8	90994951	90995147
NBN_EX2_t2	chr8	90994896	90995067
NBN_EX3_t1	chr8	90993644	90993809
NBN_EX3_t2	chr8	90993553	90993744
NBN_EX4_t1	chr8	90993024	90993175
NBN_EX4_t2	chr8	90992902	90993100
NBN_EX5_t1	chr8	90990451	90990650
NBN_EX5_t2	chr8	90990354	90990550
NBN_EX6_t1	chr8	90983400	90983598
NBN_EX6_t2	chr8	90983302	90983486
NBN_EX7_t1	chr8	90982634	90982808
NBN_EX7_t2	chr8	90982539	90982736
NBN_EX8_t1	chr8	90976650	90976820

NBN_EX8_t2	chr8	90976533	90976729
NBN_EX9_t1	chr8	90970923	90971121
NBN_EX10_t1	chr8	90967624	90967823
NBN_EX10_t2	chr8	90967541	90967719
NBN_EX10_t3	chr8	90967421	90967616
NBN_EX11_t1	chr8	90965771	90965969
NBN_EX11_t2	chr8	90965650	90965845
NBN_EX11_t3	chr8	90965531	90965724
NBN-EX11_t4	chr8	90965415	90965606
NBN_EX12_t1	chr8	90960004	90960171
NBN_EX13_t1	chr8	90958416	90958598
NBN_EX13_t2	chr8	90958309	90958508
NBN_EX14_t1	chr8	90955456	90955651
NBN_EX15_t1	chr8	90949219	90949342
NBN_EX16_t1	chr8	90947748	90947873

BARD1

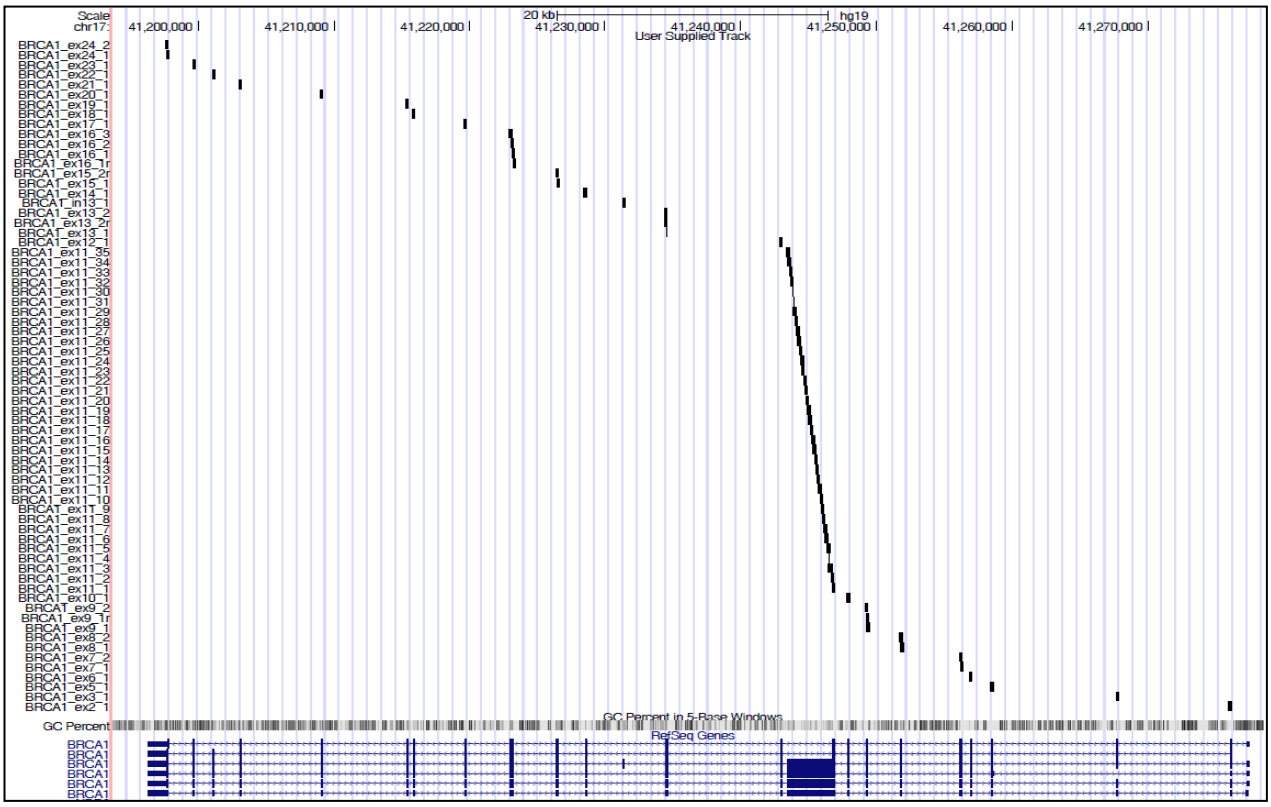
Amplicon	chr	Start	End
BARD1_EX1_t1	chr2	215674121	215674320
BARD1_EX2_t1	chr2	215661801	215661902
BARD1_EX3_t1	chr2	215657067	215657218
BARD1_EX4_ta1	chr2	215646097	215646293
BARD1_EX4_t1	chr2	215646014	215646164
BARD1_EX4_t2	chr2	215645889	215646084
BARD1_EX4_t3	chr2	215645829	215646018
BARD1_EX4_t4	chr2	215645744	215645925
BARD1_EX4_t5	chr2	215645683	215645849
BARD1_EX4_t6	chr2	215645617	215645781
BARD1_EX4_t7	chr2	215645538	215645733
BARD1_EX4_t8	chr2	215645464	215645636
BARD1_EX4_t9	chr2	215645411	215645578
BARD_ EX4_t10	chr2	215645346	215645522
BARD1_EX4_t11	chr2	215645243	215645428
BARD1_EX4_t12	chr2	215645150	215645320
BARD1_EX5_t1	chr2	215633927	215634126
BARD1_EX6_t1	chr2	215632220	215632418
BARD_ EX6_t2	chr2	215632095	215632291
BARD1_EX7_t1	chr2	215617142	215617328
BARD_ EX8_t1	chr2	215610423	215610622
BARD1_EX9_t1	chr2	215609757	215609929
BARD1_EX10_t1	chr2	215595092	215595278
BARD1_EX11_t1	chr2	215593566	215593765
BARD1_EX11_t2	chr2	215593482	215593681
BARD1_EX11_t3	chr2	215593354	215593553

Appendix X

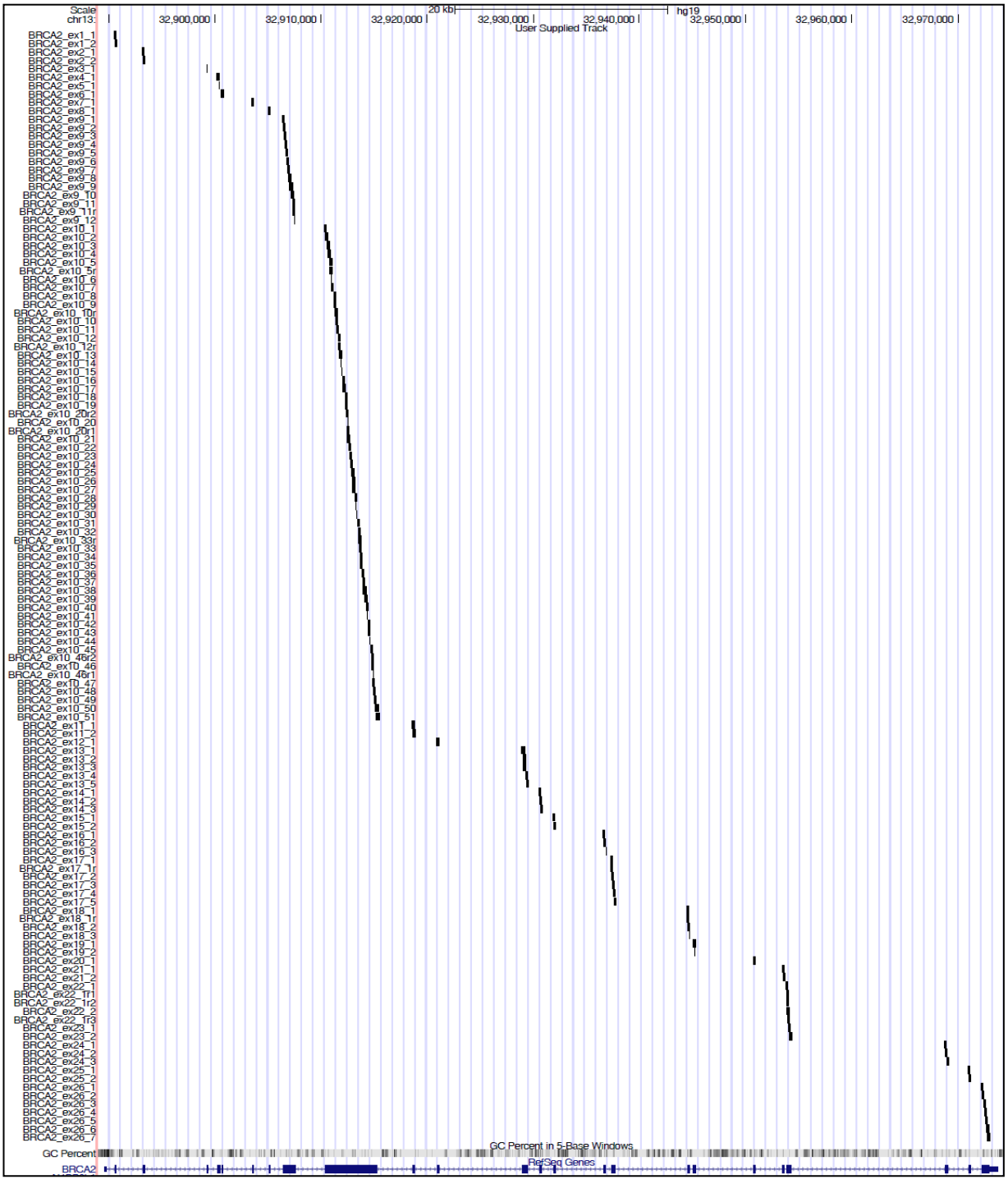
Amplicon maps of target regions for additional genes in 9-gene study

The following diagrams are amplicon maps of target regions for each gene

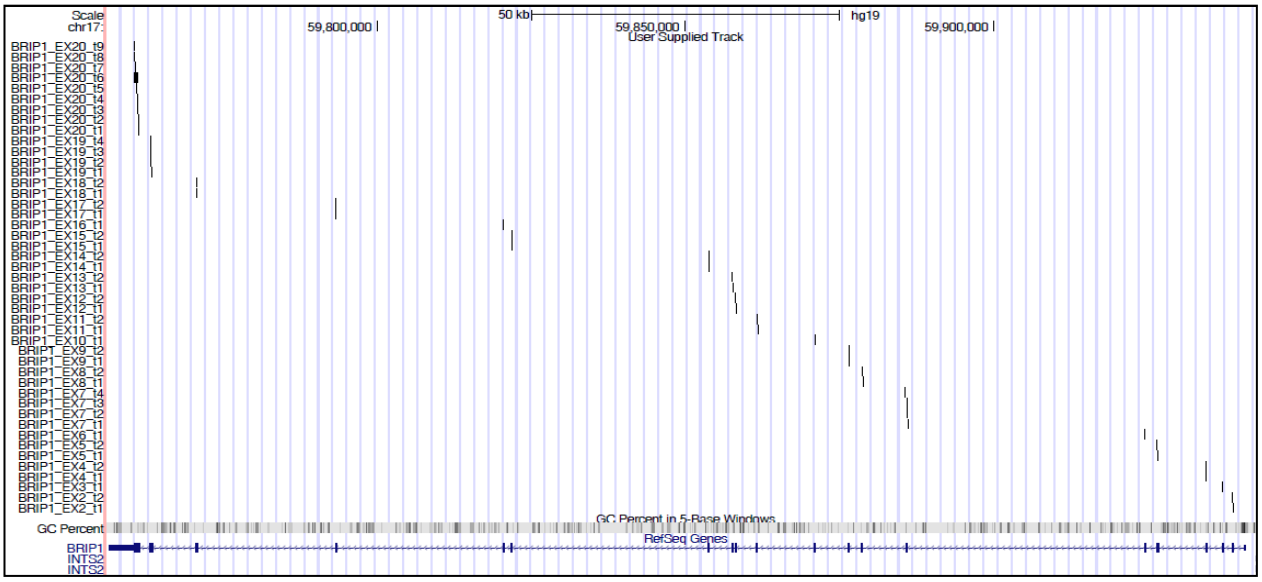
BRCA1



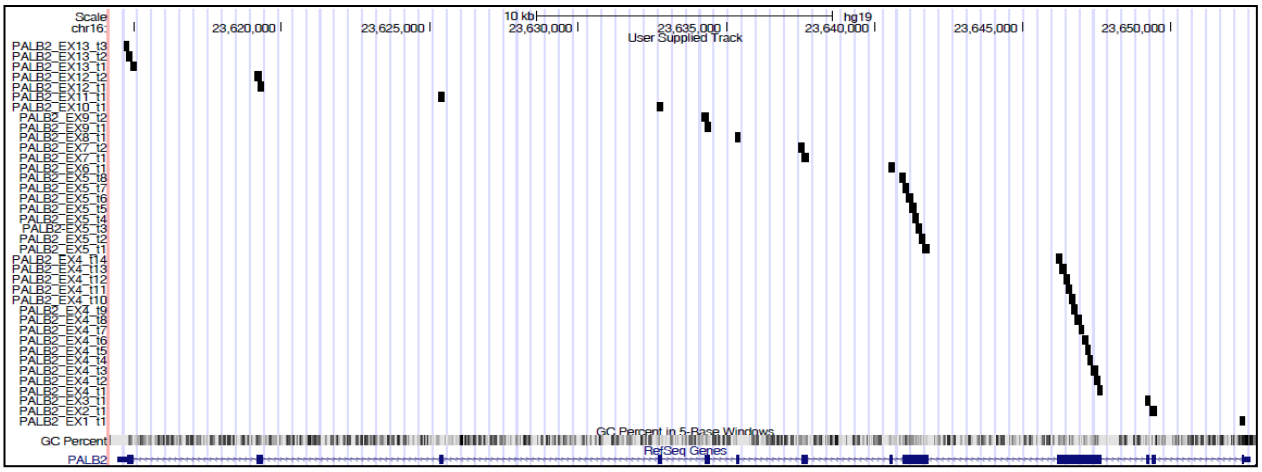
BRCA2


Scale
chr13:
20 kb
hg19
32,900,000
32,910,000
32,920,000
32,930,000
32,940,000
32,950,000
32,960,000
32,970,000
User Supplied Track
GC Percent in 5-Base Windows
GC Percent
RefSeq Genes
BRCA2

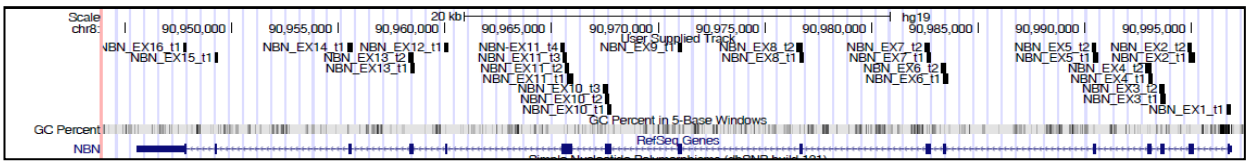
BRIP1



PALB2



NBN



BARD1